7-19-2016

# Household Water Filter Use Characterization in Rural Rwanda: Signal Interpretation, Development and Validation

Sarita Lucia Tellez Sanchez
*Portland State University*

# Household Water Filter Use Characterization in Rural Rwanda:

# Signal Interpretation, Development and Validation

by

Sarita Lucia Tellez Sanchez

A thesis submitted in partial fulfillment of the
requirements for the degree of

Master of Science
in
Electrical and Computer Engineering

Thesis Committee:
Evan Thomas, Chair
James McNames
Martin Siderius

Portland State University
2016

## Abstract

Access to safe drinking water is an important health factor in many developing countries. Studies have shown that unsafe drinking water and poor sanitation practices leads to diarrheal disease, which is one of the leading causes of death of children under five in developing countries [1]. Provision and proper use of household water filters have been shown to effectively improve health [2][3].

This thesis is focused on the refinement and validation of algorithms for data collected from pressure transducer sensors that are used in household water filters (the Vestergaard Frandsen LifeStraw Family 2.0) deployed in Rwanda by the social enterprise DelAgua Health. Statistical and signal processing techniques were used to detect the use of the LifeStraw water filters and to estimate the amount of water filtered at the time of usage. An algorithm developed by Dr. Carson Wick at Georgia Institute of Technology was the baseline for the analysis of the data. The algorithm was then refined based on data collected in the SweetLab at Portland State University, which was then applied to field data.

Laboratory results indicated that the mean error of the improved algorithm is 11.5% as compared with the baseline algorithm mean error of 39%. The validation of the algorithm with field data yielded a mean error of 5%. Errors may be attributed to real-world behavior of the water filter, electronic noise, ambient temperature, and variations in the approximation made to the field data. This work also presents some consideration of the algorithm applied to soft-sided water backpacks.

**Acknowledgements**

I would like to sincerely thank my advisor, Dr. Evan Thomas, for giving me the opportunity to work on this project and for his support, patience and encouragement during the past year.

I would also like to thank my committee members, Dr. James McNames and Dr. Martin Siderius, for serving in my defense committee, and Dr. Carson Wick of Georgia Institute of Technology.

Specially, I would like to thank my husband, my parents, sibling and friends for their love, support and strength throughout this journey.

## Table of Contents

**List of Tables**

**List of Figures**

# 1. INTRODUCTION

More than a billion people in the world continue to lack access to safe drinking water and sanitation services. Studies have shown that unsafe drinking water along with poor hygiene practices is the consequence to numerous deaths and diseases, such as diarrhea, among children under the age of five [1][7].

Interventions addressing the health impacts of improved sanitation practices and clean water are often performed through public health interventions. The impact of these programs is often measured via household-to-household surveys. These surveys are known to have limitations due to overestimation of adoption rates [3][7].

The need to acquire objective data for development programs in developing countries has brought the attention to the use of sensors as a tool to overcome the challenges of household-to-household surveys. Sensors have demonstrated to be beneficial to monitor and evaluate the successful rate of these development programs [8].

The SweetLab at Portland State University developed a cellular reporting instrumentation system, with the aim of providing objective data on the use of household health interventions in developing countries [7]. Specifically, the instrumentation system improves monitoring of water filters, water-carrying backpacks, cook stoves, and many other devices.

This thesis is focused on the refinement and validation of algorithms for data collected from pressure transducer sensors that are used in household water filters that are deployed in Rwanda. Additionally, this work also presents some consideration of the

algorithm applied to soft-sided water backpacks. An algorithm developed by Dr. Carson Wick of Georgia Institute of Technology was the departure point for this work.

This chapter discusses an overview of the sensor technology in Section 1.1. Section 1.2 describes applications of this technology: Water filter, and water-carrying backpack. The objectives and significance of this thesis are discussed in Section 1.3. Section 1.4 describes the contribution made in this thesis.

### 1.1 Sensor Technology Overview

The SweetLab at Portland State University developed a remotely reporting cellular sensor technology to monitor performance of water filters, cook stoves, and more (Figure 1). The technology consists of an integrated system that includes commercially available front-end sensors (a Honeywell digital 1-psi water pressure transducer), processing hardware, cellular network radio and power supply incorporated into a watertight enclosure. The system logs data locally and reports to the cloud on a user-configured schedule to minimize the power consumption.


(a) (b)

**Figure 1: Sensor used in water filters and water carrying backpacks. a) Watertight enclosure b) sensor.**

The cloud provides storage of data online that has multiple benefits such as storage capacity, back-up facilities, easy accessibility, and low cost. There are many providers that offer this service such as Amazon EC2, Microsoft, Google and others.

After the system reports to the Internet cloud, PHP protocols are used to process the raw data that will later be stored in a MySQL database. As a final stage, R scripts are developed to analyze the data stored in MySQL through the application of signal processing and statistical analysis [7].

The raw data of the water filter application is characterized to be non-uniformly sampled; this is due to the fact that it is an event-triggered application. The analysis of data involves laboratory testing to make claims about data from the field, re-sampling of the data to facilitate analysis, detecting events during which the LifeStraw was used, and estimating amounts of water filtered or carried by these devices respectively.

### 1.2 Technology Application: Water Filter and Water Carrying Backpack

- **LifeStraw Description**

The Vestergaard Frandsen LifeStraw Family 2.0 (Figure 2) is a tabletop water filter designed for households in developing countries. Microbiological contaminated water is poured through an 80-micron textile pre-filter into a six-liter storage tank. The water is then filtered at a flow rate of approximately 1.2 liters per hour and passed into a safe storage tank with a capacity of 5.5 liters, where it can be dispensed using a plastic water tap. The filter can be backwashed by pressing the backwash lever and the dirty water is placed in the backwash tank. These filters are able to treat 20,000 liters of water; and has

been reported that the can provide enough water for a family of five for at least three years [2][3].

The sensor for this application includes a Honeywell digital 1-psi differential pressure transducer located in the bottom reservoir of the LifeStraw. This sensor records pressure data that is associated with the filling of the bottom storage tank as the water is filtered.

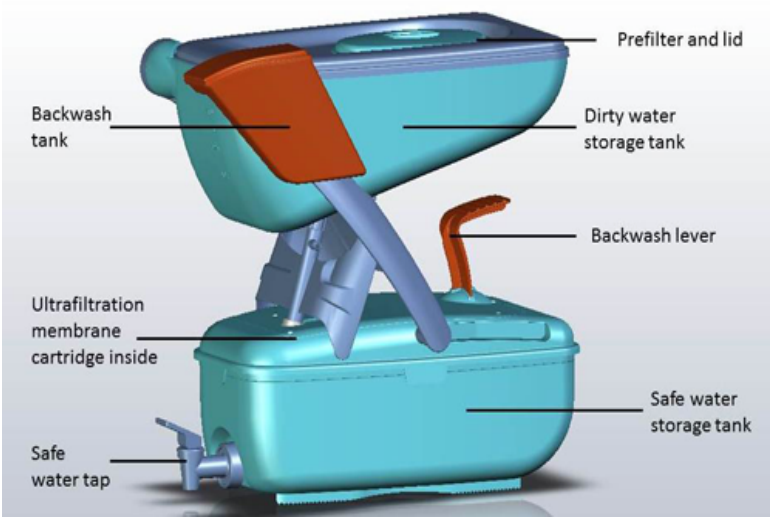



**Figure 2: Vestergaard Frandsen LifeStraw Family 2.0 [2]. Tabletop water filter designed for households in developing countries.**

- **PackH2O Description**

In water-stressed regions, women and children walk an average of 3.5 miles daily to collect water to bring to their homes. They do such a task by carrying jerry cans or old buckets on their heads. These methods are physically demanding and can lead to health problems, such as spinal pain or other joint problems [4]. The PackH2O, Figure 3, is a clean and ergonomic alternative for water collection practices. The backpack is a soft-goods bag made of Teflon enclosed in canvas with a volume capacity of 20 liters. It has a wide mouth to facilitate filling, tapered sides to minimize water leakage, and pull straps

to facilitate carrying [4].“It displaces weight on shoulders with less pressure on head, hands and neck when user is in transit”. It also has a “removable liner that is better suited to clean than a bottle neck jerry can” [5].

For data collection, a sensor-integrated system is located inside the backpack. This sensor detects movement of the backpack and reports pressure readings and location via integrated GPS.

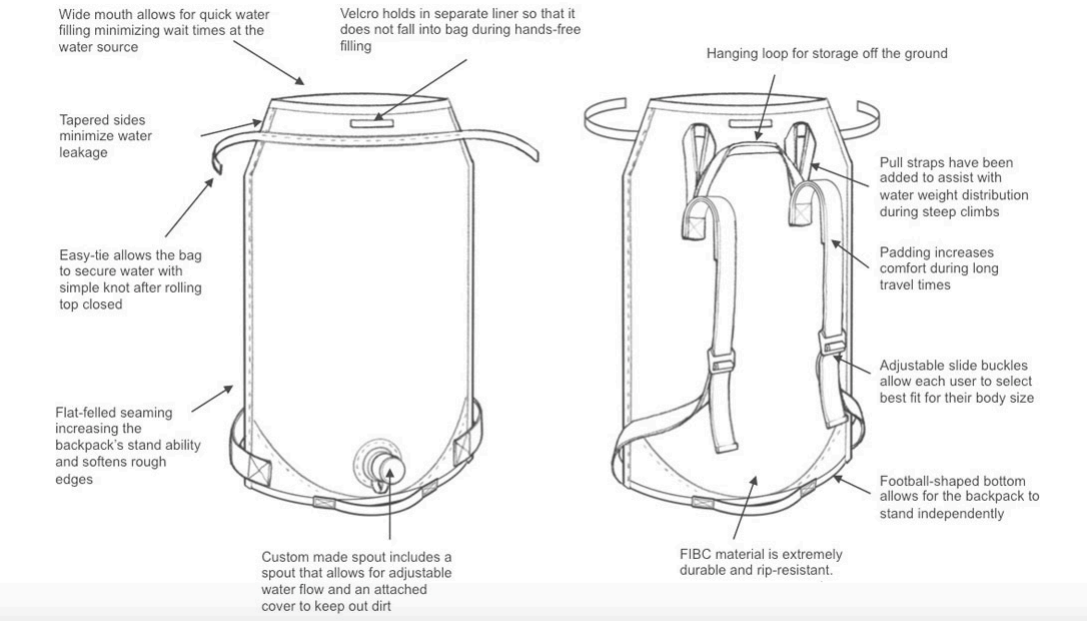



**Figure 3: PackH2O design description [4].**

### 1.3 Objectives and Significance

This research identified objectives to refine and validate algorithms for data collected from pressure sensors that are used in water filters (the Vestergaard Frandsen LifeStraw Family 2.0) deployed in Rwanda by the social enterprise DelAgua Health. This work builds upon the algorithm developed by Dr. Carson Wick at the Georgia Institute of Technology.

These objectives include:

- Detect filling events.
- Estimate volume of filling events.
- Validate analysis with laboratory data.
- Analyze and validate field data.

The success of effectively refining the algorithm and analyzing data will allow demonstrating more accurately the success of the cellular reporting instrumentation system. This is because the refined algorithm tracks more precisely the behavior of the data.

## 1.4 Contributions

In this thesis, we studied the use of pressure transducer installed in household water filters. Primarily, we addressed the signal analysis and validation of algorithms developed to evaluate this data.

The analysis was based upon an algorithm made by Dr. Carson Wick and modifications in the algorithm were determined by the accuracy on detecting events and estimating volume. This modification were based on the hypothesis that the error in the algorithm came from error calculating the slope of the event, start and stop time of the event, or the digital units per liter, i.e. the expected change of the pressure reading per liter.

The contributions are:

- Field data was manually reviewed to identify drawbacks in the baseline algorithm, and laboratory data was collected to make the necessary modifications to the algorithm accordingly to flaws identified in the analysis.
- The compensation of the temperature was taken into account by collecting data from a LifeStraw exposed to temperature changes and using regression techniques to minimize the effect of the heating and cooling cycles.
- The calculation of the slope was improved by modifying parameters and thresholds that allow easier and better estimation of the slope.
- The estimation of volume was improved by redefining the relationship between the digital units of pressure and a litter of water. Also, 0.5 liters of water were added to final result because it was found to be the minimum detection limit.

- Validation of the algorithm using a second round of datasets of LifeStraws deployed in Rwanda.
- Precision and recall analysis to evaluate the accuracy of the algorithm.

## 2. LITERATURE REVIEW

Literature addressing the analysis of data from pressure transducers installed in household water filters in developing countries has not been found. Therefore, this chapter summarizes relevant studies with a focus on temperature compensation of pressure transducers, and also publications related with other electronic sensors installed for development programs. Section 2.1 summarizes some of the literature found, and Section 2.2 addresses the relationship and relevance of these studies with the analysis of data from pressure transducers installed in household water filters.

### 2.1 Summary of Previous Work

Studies have shown that sensors have become promising tools to be used in development programs to gather unbiased and more precise data [8]. This technology, and the need to overcome challenges of previous methods to evaluate development programs, has led to an increase in research and expanding the range of what can be measured.

*Mercado et al*. proposed the use of sensors to facilitate collection of data. In their study, they deployed temperature dataloggers as Stove Use Monitors (SUMs) and implemented an algorithm to obtain the daily usage of the stove. Their project was implemented in 80 rural households for 32 months in Guatemala using a chimney-cookstove [16]. For signal analysis, *Mercado et al.* implemented a peak selection algorithm based on the instantaneous derivatives and the statistical behavior of the stove

and the surrounding temperature signals. The data for this application is temperature data, with a 20 min sampling rate.

The signal peaks were detected using data analysis and graphing software (OriginLab). Then the positive peaks were filtered from peaks related to indoor temperatures, using threshold slope values. The threshold slopes (S0- and S0+) were obtained from the days in which the chimney cook stove was not in use, the 1$^{st}$ and 99$^{th}$ percent of those derivatives were chosen to be S0- and S0+ respectively. Fueling events were considered as the positive peaks when the thresholds from the positive and negative slope was exceed. Fueling events separated by less than 2 hours were clustered and counted as a single cooking event.

*Rostapshova et al.* propose the use of pressure loggers to measure intermittent urban water supply in Tanzania [8]. Each device connects to the end of a yard tap and logs pressure continuously (every 10 minutes) for months. Collected data was used to analyze the number and duration of outages, and average hours of water service per day.

Other relevant work influencing the scope of this study is the effect that temperature has on the performance of pressure sensors. In this issue, *Palmer* proposed an interpolation algorithm method to compensate the temperature influence in pressure sensors; his algorithm is based on the Memscap SP 82 pressure sensor [13]. The interpolation algorithm is based upon a set of functions that represent the pressure characteristics over temperature. These functions need to be an accurate fit to the data to allow the output signal of the sensor to be described. The functions used in this work were 5$^{th}$ order polynomials.

The idea of his algorithm is that the temperature and sensor signal be the input of a single equation. The temperature is used to calculate where on the x-axis the range of temperature is found, and the signal input (mV) defines where on the y-axis the value should be. The solution of the algorithm lies upon the intersection of the x and y-axis.

The interpolation algorithm is then developed by finding the boundaries around the area of interest i.e. narrowing the values from the upper and lower boundaries until they match. Each of the upper and lower values is assigned a buffer value. If the input signal value falls in either the upper or lower boundary, a solution is found.

*Watras et al.* investigated the effect of temperature on fluorescence sensors [17]. The experiments that they developed show that CDOM (chromophoric dissolved organic matter) fluorescence intensity decreased as ambient water temperature increased. A temperature compensation equation was derived and applied to field data. This removed the effect of multi-day trends in water temperature. The equation is represented as:

$$CDOM_r = \frac{CDOM_m}{1 + \rho(T_m - T_r)}$$

Where T is temperature (°C), ρ is the temperature-specific coefficient of fluorescence ($°C^{-1}$), m is the slope, and C represents the intercept of the CDOM regression. The subscripts r and m stand for the reference and measured values. The intensity had linear dependency on the temperature and the effect was reversible during sequential heating-cooling cycles. The temperature compensation removed the effect of gradual cooling, eliminating an upward trend in CDOM.

## 2.2 Relationship

The studies summarized in the previous section described different projects involving the use of sensors in development programs, and temperature compensation algorithms.

The document from *Mercado et al.* describes a method to gather the data related to the daily uses of a cookstove. The data from this project is more straightforward than the data collected with the pressure transducer used in the water filter. In this project, the events are represented as the temperature increment that surpasses a certain threshold (Figure 4), and the analysis is focused on the selection of those temperature peaks. In comparison, the data collected with the pressure transducer is highly influenced by external factors, which impacted the development of the analysis. However, as with the cookstove project from *Mercado et al.*, thresholds were also used to choose the slope data that represent a filtering event.

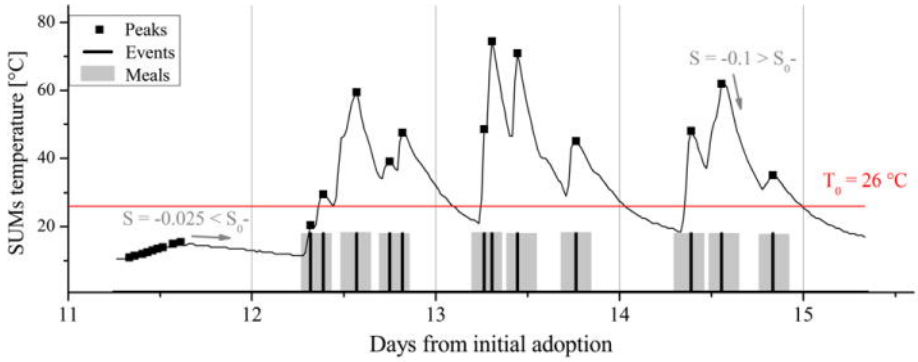



**Figure 4:Analysis of data from Mercado *et al.* project. The signal represents temperature data from a chimney cookstove.**

The temperature compensation algorithm served as a starting point to analyze the effect of temperature on the pressure data. Palmer found a set of equations that he

obtained by testing the pressure sensor exposed at different temperatures ranges. Then he developed an interpolation algorithm that is based upon the known input signal and the temperature, these allow obtaining a value close to the expected output. Even though Palmer's study looks promising, as a way of reducing the temperature effect in the pressure data, there is not a clear relationship among the pressure data collected from the water filter, the volume of water added, and the temperature that allows Palmer's approach to be used.

Watras *et al.* study was based on the effect of temperature on fluorescence sensors in freshwater. In their study they found that there is some degree of hysteresis in the heating and cooling cycles, i.e. when the temperature increases the intensity of the fluorescence decreases and when the temperature decreases the intensity of the fluorescence increases. Their study has some degree of similarity with the effect of temperature on the raw pressure data, in the sense that there is also some degree of hysteresis affecting the pressure sensor. This study served as a base to reduce the effect of temperature in the raw pressure data.

# 3. MATERIALS AND METHODS

The goal of this chapter is to provide a description of the database and the procedure taken to collect the data. Section 3.1 described the database from which the data was retrieved, and Section 3.2 describes the data collection protocol. The data collection section describes the data collected from the water filter and gives a brief description of the data collected from the backpack and the challenges encounter with it.

## 3.1 Description of Database

The data used in this work is stored in an Amazon EC2 hosted MySQL database, which contains the datasets from 174 different LifeStraw and 37 PackH2O units deployed in Rwanda and Haiti respectively. The LifeStraw and PackH2O databases include raw data corresponding to the time in which the data was reported, pressure, temperature, battery life, and the signal strength (RSSI). Additionally, the PackH2O database contains accelerometer data and GPS coordinates.

The sensor platform for the LifeStraw logs data every 5 minutes and transmits the data to the Internet cloud every 24 hours. For the PackH2O, logging occurs every 5 minutes after the sensor is triggered and the user has walked with the backpack for more than 5 minutes, the transmission of data to the cloud occurs once every 24 hours. There were some data gaps, which were attributed to cellular connectivity dropouts. Some other data gaps are attributed to depleted batteries.

The data collected from these sensors represents the relationship between raw pressure data and time. In this data the water filtration events are represented as a linear

increase of pressure over time. The range of the raw pressure data is the correlation of the SPI data bus input from the Honeywell 1-psi differential pressure transducers.

## 3.2 Data Collection

### 3.2.1 Water filter: LifeStraw

Laboratory experiments were developed to assist in the collection of data from the pressure sensors. This data then was used to refine the algorithm. Prior starting the experiments, analysis of field data was observed and manually reviewed to identify the points in which the algorithm was giving false results. False results include: detection of false events (false positives), no detection of events (false negatives), and wrong estimation of water volume (error precision).

The manual review consisted of a meticulous observation of the raw data to identify a continuous diagonal line, which describes the increment of pressure over time, and a roughly approximation of the volume of the water. A detail explanation of the manual review is shown in Appendix A.

Additionally, results of the analysis of laboratory data with the baseline code are included in this report to highlight the need for validation and refinement of the data-processing algorithm. These results are presented in Table 1. Table 1 shows an example of data collected in the laboratory and analyzed with the baseline code. In the table it is noticeable the discrepancies from the real/expected volume and the baseline algorithm; it has an average error of 41%, one false event, and one not detected event.

**Table 1: Comparison Between Manual Review and Baseline Algorithm Output**

| Real Volume | Digital Pressure Reading | Recorded Time of Event | Volume from Baseline Algorithm | Time from Baseline Algorithm |
|---|---|---|---|---|
| **4.5** | 9124.36 | 4/13/15 17:55 | **3.30** | 4/13/15 18:11 |
| **--** | -- | -- | **1.30** | 4/14/15 20:22 |
| **4** | 8675.46 | 4/14/15 20:07 | **0.69** | 4/14/15 21:56 |
| **3** | 8538.44 | 4/16/15 15:12 | **2.01** | 4/16/15 15:21 |
| **2** | 8280.29 | 4/20/15 17:38 | **1.05** | 4/20/15 17:51 |
| **1** | 8063.86 | 4/21/15 17:02 | **--** | -- |
| **5** | 8791.16 | 4/22/15 18:10 | **3.18** | 4/22/15 18:21 |
| **4** | 8588.70 | 4/23/15 18:53 | **2.68** | 4/23/15 18:56 |
| **3** | 8408.28 | 4/27/15 16:30 | **1.83** | 4/27/15 16:36 |
| **2** | 8246.58 | 4/28/15 19:26 | **1.31** | 4/28/15 19:26 |
| **1** | 8074.42 | 5/1/15 15:33 | **0.43** | 5/1/15 15:46 |
| **5** | 8786.51 | 5/5/15 19:50 | **3.29** | 5/5/15 20:01 |
| **3.8** | 8674.14 | 5/7/15 18:58 | **2.91** | 5/7/15 19:01 |
| **2.8** | 8364.14 | 5/11/15 17:47 | **1.65** | 5/11/15 17:51 |

The complete test plan developed to collect data and validate the algorithm is presented in Appendix B. The thought process behind each of these experiments is explained in the following paragraphs.

*Test ID 1:Pressure units of know known volume of water filtered*

Analysis of the field data (Appendix A) indicated that the baseline algorithm does not generate consistent volume event estimates for comparable raw pressure readings. As an example, Table 2 presents a summary of the results given by the baseline algorithm. It shows some examples of the volume, the corresponding pressure reading and intercept.

From the table it can be seen that there is a difference in volume results from comparable pressure reading. As an example LSF20_A62F3A presents pressure readings greater than 8900 digital units with volume values of about 0.52 liters, while

LSF20_A60317 with 8893 raw pressure units gave 4 liters of water. The differences on these results seem to be due to the difference on the intercept. The intercept is the lower pressure value for the event, in these cases it can represent that the algorithm picked up noise as a real event or that there are more that one event in a day. The first assumption seems more realistic as the volume estimated is very small. As mentioned earlier, the complete analysis and manual review is presented in Appendix A.

**Table 2: Volume and Pressure of Different LifeStraws Units**

| LSF20_A62F3A | | LS20_A60317 | | LSF20_A5DEB2 | |
|---|---|---|---|---|---|
| **Vol (L)** | **Pres/Int** | **Vol (L)** | **Pres/Int** | **Vol (L)** | **Pres/Int** |
| 0.68 | 8409.60/ 8255.79 | 0.91 | 8362.29/ 8155.89 | 1.68 | 8740.12/ 8360.63 |
| 1.20 | 8691.12/ 8421.20 | 1.61 | 8748.64/ 8384.78 | 1.36 | 8735.76/ 8429.25 |
| 0.15 | 8576.44/ 8542.93 | 0.54 | 8621.78/ 8501.46 | 0.12 | 8746.32/ 8719.332 |
| 0.18 | 8634.05/ 8593.92 | 3.17 | 8723.89/ 8008.84 | 0.58 | 8510.96/ 8379.32 |
| 1.23 | 8632.69/ 8356.50 | 3.53 | 8740.98/ 7946.60 | 1.56 | 8686.40/ 8334.28 |
| 0.52 | 8907.23/ 8790.31 | 4.01 | 8893.77/ 7989.7 | 0.13 | 8453.39/ 8423.82 |
| 2.17 | 8543.57/ 8054.95 | 0.72 | 8216.69/ 8053.70 | 0.17 | 8437.18/ 8398.75 |
| 0.71 | 8839.69/ 8679.10 | 0.37 | 8854.43 /8769.89 | 0.36 | 8429.47/ 8347.08 |
| 0.40 | 8769.79/ 7900.33 | 3.89 | 8864.57/ 7988.47 | 0.29 | 8351.15/ 8284.17 |
| 0.68 | 8409.60/ 8135.85 | 0.90 | 8284.45/ 8080.64 | | |

In this manner, test 1 is designed to obtain a correlation estimate of the relationship between the raw and known amounts of water filtered by the LifeStraw, and examine whether the pressure readings from different LifeStraw sensors were approximately the same for known volumes of water.

The overall test involved adding known amounts of water to the upper storage tank of the LifeStraw for filtering. After the data was reported register the corresponding pressure reading and use the baseline algorithm for analysis.

This test allowed an improved understanding of the relationship between the raw pressure units and liters of water, which will also benefit for the manual review needed on more field data.

*Test ID 2: Temperature effect on data*

Temperature has been an issue when dealing with pressure sensors. It has been reported that LifeStraws in Rwanda are exposed to temperature changes of about 15 to 32 degrees Celsius [11]. Figure 5 illustrates an example of the temperature signals recorded from the water filters.

The variations in temperature lead to both discrepancies in pressure readings and noise in the data. Therefore, the aim of test 2 is to verify the effect of temperature on the pressure data by exposing the LifeStraw to temperature changes throughout the day. This was accomplished by placing the LifeStraw unit outside, adding known amounts of water to the upper storage tank for filtering and, observing the behavior of the data under these conditions. Test 2 also allowed verification of whether the algorithm was able to accurately detect events in noisy environments.

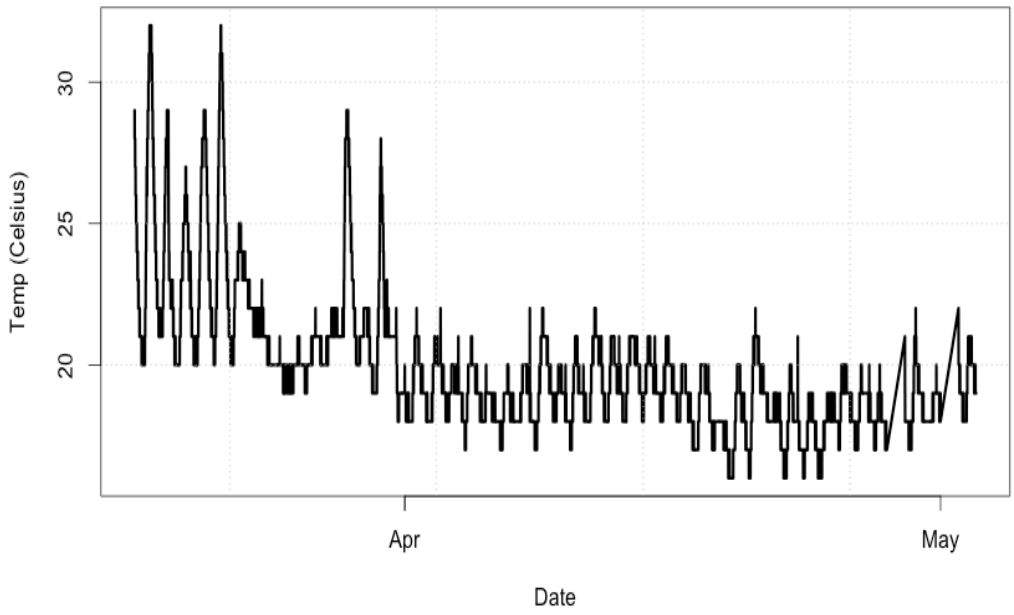



**Figure 5: Temperature signals recorded with the sensor used in a LifeStraw deployed in Rwanda. The figure shows temperature variations over the curse of three months. It is visible that around the month of March the temperatures vary from 20℃ to 33℃, and April and May temperatures were at about 15 to 22℃.**

*Test ID 3: Multiple events per day*

Typical use of the water filter was anticipated to be at least several times per week, and up to several times per day. Observation and manual review of the field data suggested that the noise observed could be misinterpreted as a real event. Figure 6 shows an example of how noisy data has been misinterpreted as real events. A red 'x' represents the detected events. Real events are considered to be a continuous diagonal line with a pressure increment of more than 200 raw units. In the figure, the baseline algorithm detected 32 events, but manual review suggests that there are 13 promising real events that are highlighted by a red number on top of the continue line.

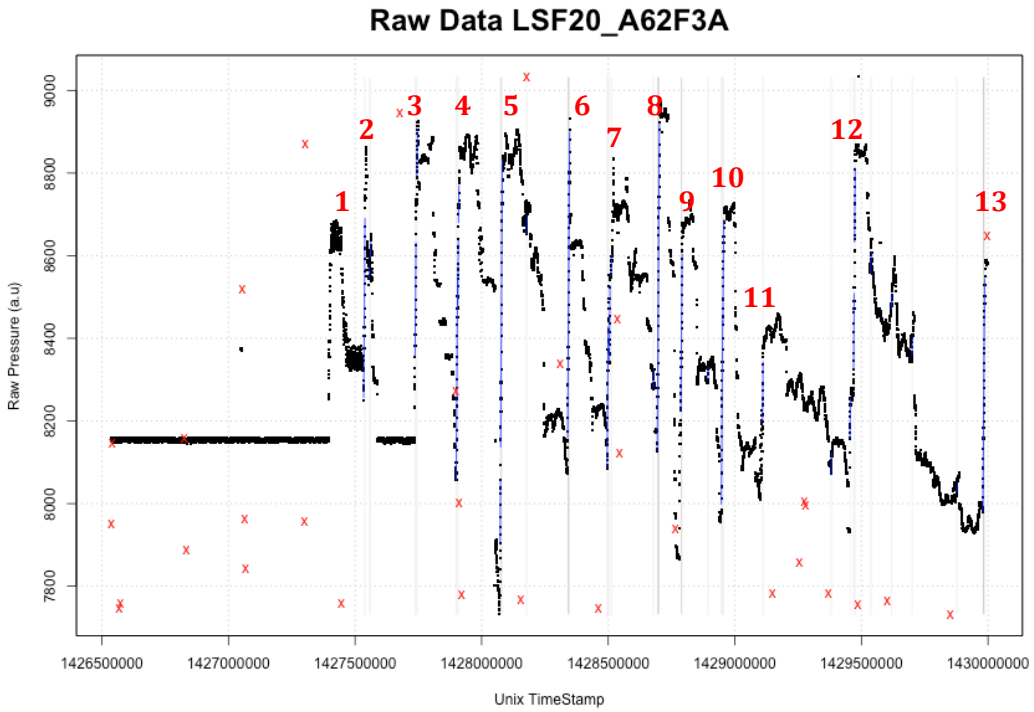



**Figure 6: Example of analysis of field data. The figure illustrates how noisy data it's been detected as real events; Real events are considered to be a continuous diagonal line. The events detected the algorithm are represented by a red 'x'. There are 13 visible real events highlighted by a red number on the top, the algorithm is picking up 32 events.**

Multiple events per day are represented as continues lines that differ slightly in the slope of the diagonal. These are assumed to be a combination of shorts events or a short and a long event[1]. False events are generally detected as short events, therefore the misinterpretation between a false and a real event.

In this manner, test 3 was designed to evaluate if the baseline algorithm is able to detect multiple events a day and real short events. This was accomplished by: filtering and draining known amounts of water throughout the day for several days, and verifying whether the baseline algorithm was able to detect the correct events.

[1] Short events are events where the amount of water filtered is two liters or less. Long events are events where the amount of filtered water is three to five liters.

### 3.2.2 Water-Carrying Backpack: PackH2O

Laboratory experiments were conducted to collect data and make assertions about the behavior of the backpack data while carrying certain amount of water.

The test conducted with the backpack included:

1. The first test conducted involved filling the backpack with different measured volumes of water, leaving it stationary for one hour or more, and then draining the water.

   After several experiments were conducted and the data reported, several issues were noticeable such as: water leakage, battery life running out within two weeks, and an irregular behavior of the data. Figure 7 illustrates an example of the data. The figure shows a quick increment of pressure representing water filling and picking up of the backpack and slow decrease of pressure implying a water leakage. Due to these issues, counter-measures had to be taken to get better performance of the backpack. One of these counter-measures was, decreasing the sensitivity level of the sensor so that the accelerometer would pick up movement when the backpack was being held and the user was walking for more than five minutes. This allowed the sensor to save battery life, which is important for users in the field. Other improvement involved, changing the bag and putting patches where there was a possible water leakage.

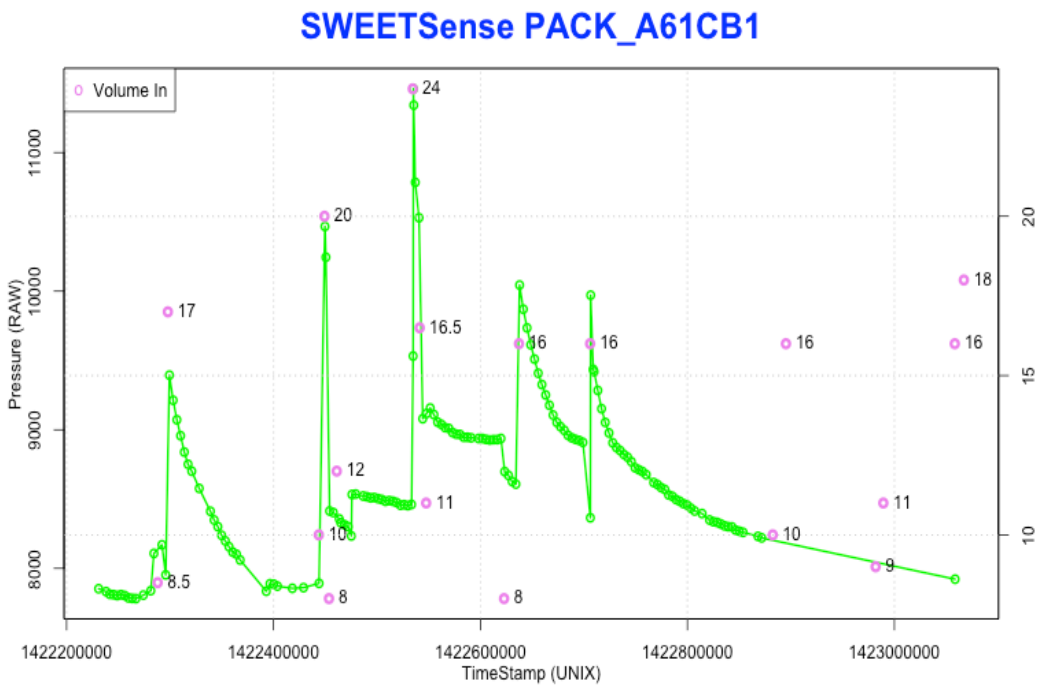



**Figure 7: Example of initial data collected from the backpack. Green points refer to raw pressure data overtime; the pink circle denotes the volume added. From the figure it can be seen that there is quick increase on pressure referring to the filling of water backpack, the slow decrease in pressure refers to water leakage.**

2. After improvements to the backpack, a second test was performed to try to understand the behavior of the data. This test was similar to the previous one with the difference that this test required the person to walk with the backpack for more than 10 minutes.

   The improvements made to the sensor allowed battery saving but it reduced the number of data points collected. Changing the bag and putting patches reduced the leaks, but did not completely stop the leakage. This challenges made difficult to understand the data, as its behavior was unpredictable. Figure 8 illustrates an example of a set of water collection events from the backpack. The figure shows data points in blue and volume collected with the pack in pink. For example, the figure shows that when four liters of water were collected with the pack the

maximum pressure value reported was 9000 raw pressure units, six liters of water reported 8800 units of pressure, and eight liters of water reported 9200 units of pressure.

The uncertainty within the data collection led to challenges in analyzing the data and understanding the behavior of the backpack. Further improvements in the design of the bag and improved battery life, might allow to obtain better data that would be within expected tolerance levels.

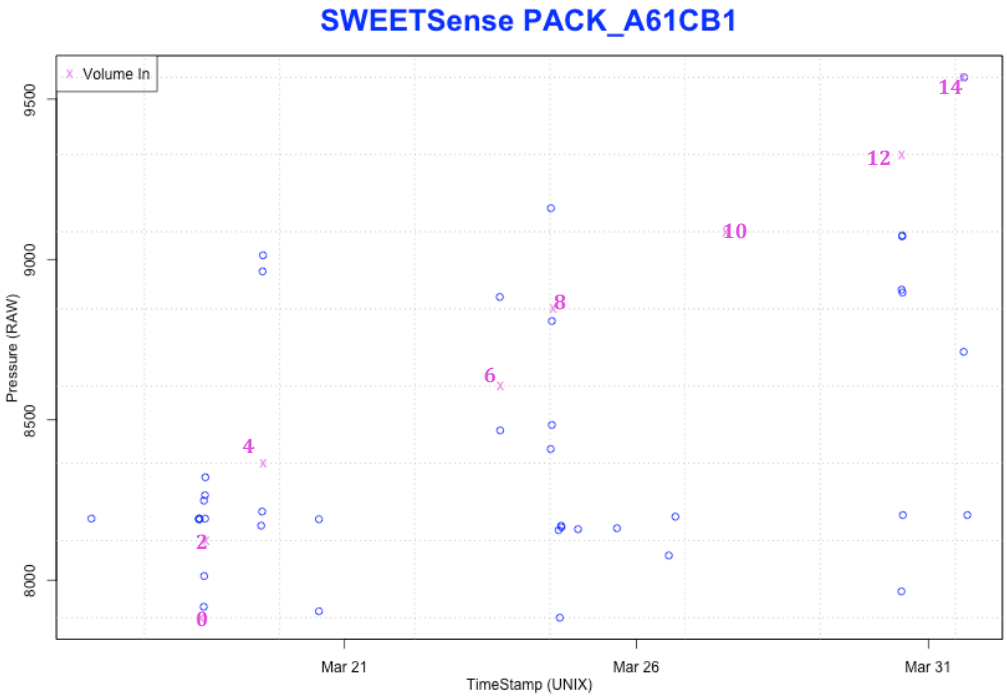



**Figure 8: Example of backpack data. The figure illustrates seven water-collected events from the backpack after improvements were made to the sensor. The raw data is shown in blue and the liters of water collected in pink. There is not a consistent behavior in the data to be able to make assumptions that allow analysis**

## 4. DATA ANALYSIS

The goal of this chapter is provide a description of the algorithm and the improvements made to it.

The LifeStraw algorithm refinement effort was based on an algorithm developed by Dr. Carson Wick at the Georgia Institute of Technology. This algorithm is focused on calculating events and usage using R (R Foundation for Statistical Computing) [12].

The algorithm involved three main parts:

- Preprocessing the raw data
- Detection of events
- Estimation of volume of water

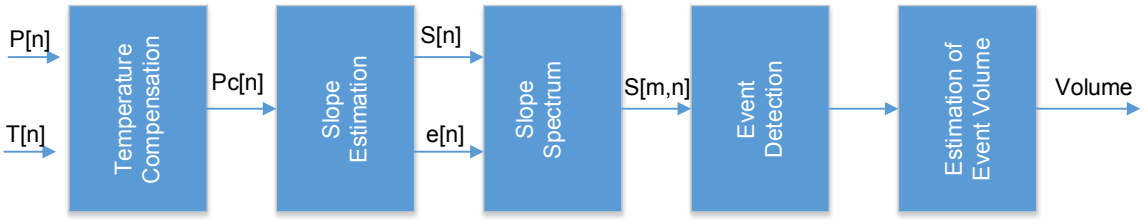



**Figure 9: Algorithm Steps**

Each of these parts was studied and adapted to improve results in the detection of events and the accuracy in the estimation of water. It is assumed that the error in the analysis could be due to wrong estimation of slope of the event, start and stop time of the event, or the digital units per liter, i.e. the expected change of the pressure reading per liter. The following sections describe the baseline algorithm and discuss the thought

process behind the adjustments and improvements made to the algorithm. A flowchart representing the algorithm is shown in Appendix C.

### 4.1 Preprocessing

The first step in developing the algorithm involves preprocessing of the raw data, which is focused on cleaning the signal to allow analysis. To accomplish this task the raw data is plotted in order to observe the sensor output characteristics. This allowed determining the need for removing duplicates of data in time, and interpolating spurious points in pressure (outside the working range).

Additionally, it was observed that daily temperature changes introduced noise in the digital pressure data. Originally, a linear temperature correction was performed to address this limitation. This was accomplished by exposing the LifeStraw and sensor to temperature variation while keeping the volume and pressure constant. These measurements were used to minimize the total pressure error in a least-squares sense. This error is defined as:

$$e_p[n] = p[n] - \hat{p}[n]$$

Where $p[n]$ is the *known* pressure and $\hat{p}[n]$ is the *corrected* pressure.

The temperature-corrected pressure is then defined as

$$\hat{p}[n] = a \cdot \tilde{p}[n] + b \cdot T[n] + c$$

Where $\tilde{p}[n]$ is the raw pressure reading, $T[n]$ is the temperature reading, and $a$, $b$, and $c$ are the fitting constants.

***Improvements:***

It was noted that the effect of the temperature was still a significant source of error even after using the compensated temperature method. The idea of compensating the temperature is to reduce the effect of the heating and cooling cycles with the aim of obtaining pressure values concentrated around their mean regardless of the temperature.

Therefore in order to obtain more accurate results, efforts were made to improving the model. To accomplish such a task, multiple pressure readings were obtained at multiple known volumes over a range of temperatures (daily temperature changes). This data was then plotted in order to observe the relationship between temperature and pressure. Figure 10 and Figure 11 show examples of the raw data and temperature of two LifeStraws sensors that were used for laboratory experiments.

Figure 10 represents data from a LifeStraw unit exposed to temperature changes. The figure shows that the baseline algorithm detected three false events, which can be attributed to the effect of the temperature. As comparison, Figure 11 shows data from a LifeStraw sitting in the laboratory. In this figure it is clearly seen that the temperature is more constant and the data is smoother than the previous figure.

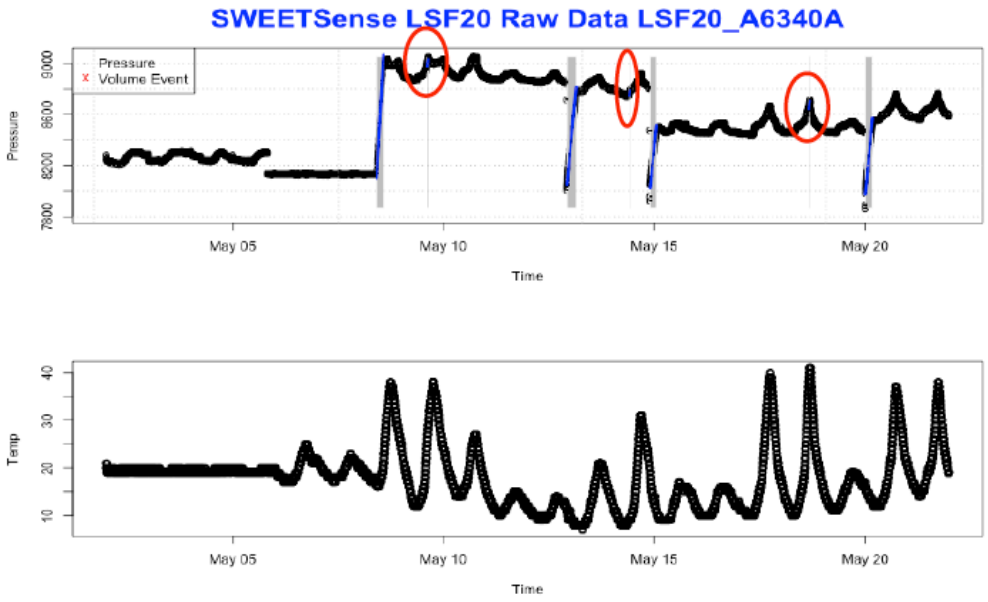



**Figure 10: Laboratory data from a LifeStraw exposed to temperature changes. Top figure shows raw pressure data over time and bottom figure temperature over time. When water was kept in the bottom basin the pressure varied accordingly to the temperature changes. The red circles indicates the three false events detected by the baseline algorithm**

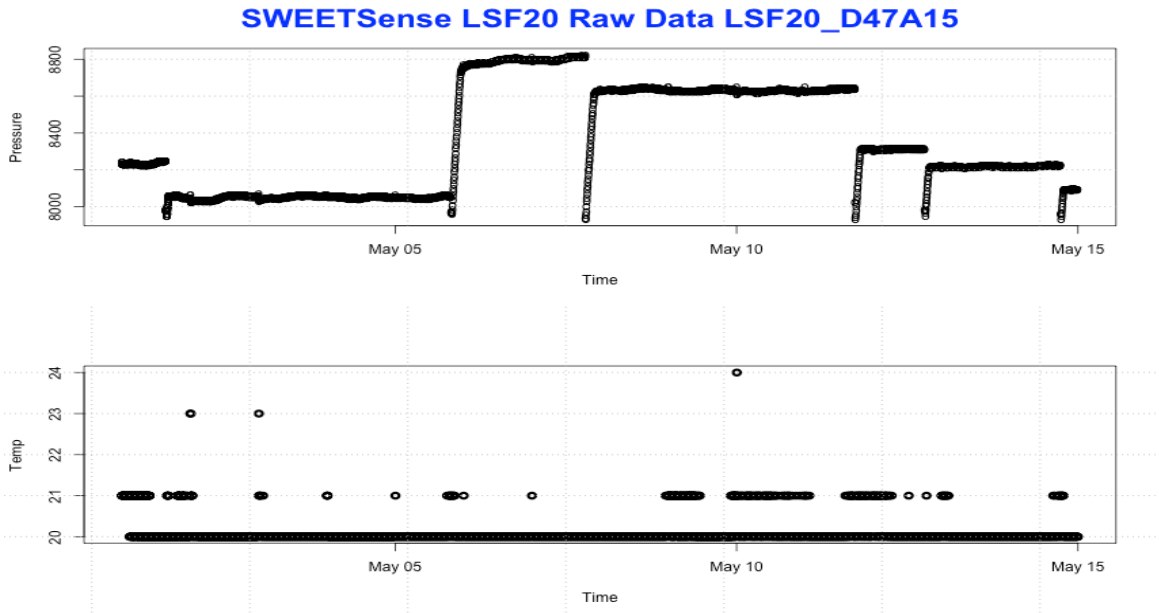



**Figure 11: Raw laboratory data not exposed to temperature changes. When water was kept in the bottom basin the pressure maintain constant over time.**

The effect of temperature on the LifeStraw data was investigated through laboratory experiments by adding known volumes of water to the bottom basin of the LifeStraw that was placed outside and exposed to temperature changes. The temperature at which the LifeStraw was exposed was reported to vary from 5 to 25 degree Celsius. Figure 12 and Figure 13 show examples of the relationship between pressure and temperature. From the figures it is noticeable that the pressure readings are positively co-related with temperature readings, i.e., it decreases as temperature decreases and increases as temperature increases.

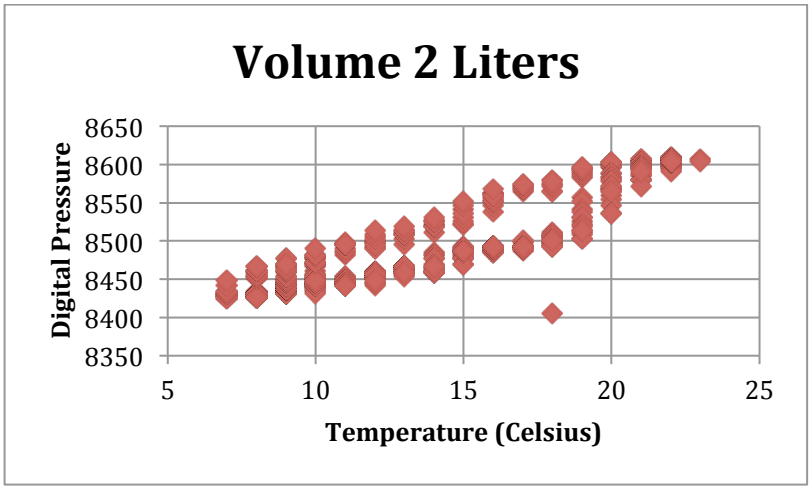



**Figure 12: Temperature effect on pressure data for two liters of water added in the bottom storage tank of the LifeStraw. The LifeStraw was exposed to daily temperature changes.**

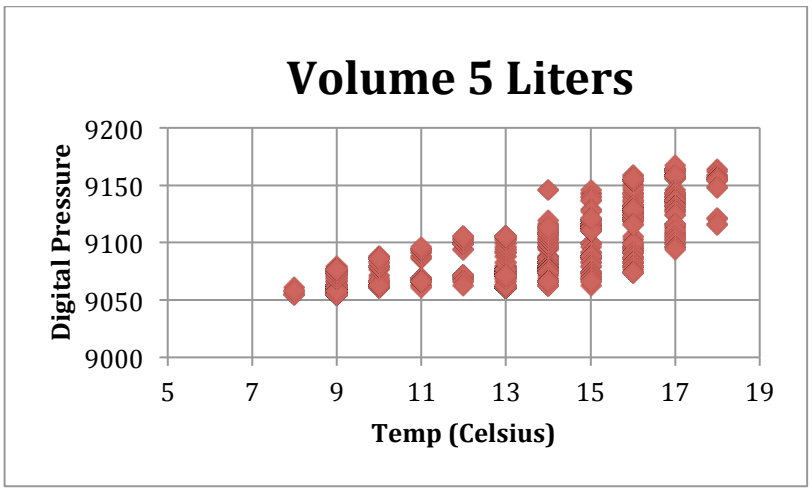



**Figure 13: Temperature effect on pressure data for five liters of water added to the bottom storage tank of the LifeStraw. The LifeStraw was exposed to daily temperature changes.**

To be able to compensate the temperature effect, the method proposed by *Watras et al*. was applied to standardize the pressure reading to a reference temperature.

$$P_c = \frac{P_m}{1 + \rho(T_m - T_c)}$$

Where $P$ represents pressure, $T$ temperature, and $\rho$ the temperature coefficient. The subscripts c and m represent the corrected and measured data respectively. The temperature coefficient was estimated by doing an iterative check of the baseline of the pressure data collected. This process was intended to improve the quality and functionality of the design.

After all the necessary improvements, the model was applied to the raw pressure data. Figure 14 shows an example of the effect of the temperature correction model on the raw pressure data for two liters of water. It can be seen that there is less variability in the pressure data regardless of the temperature at which the LifeStraw is exposed.

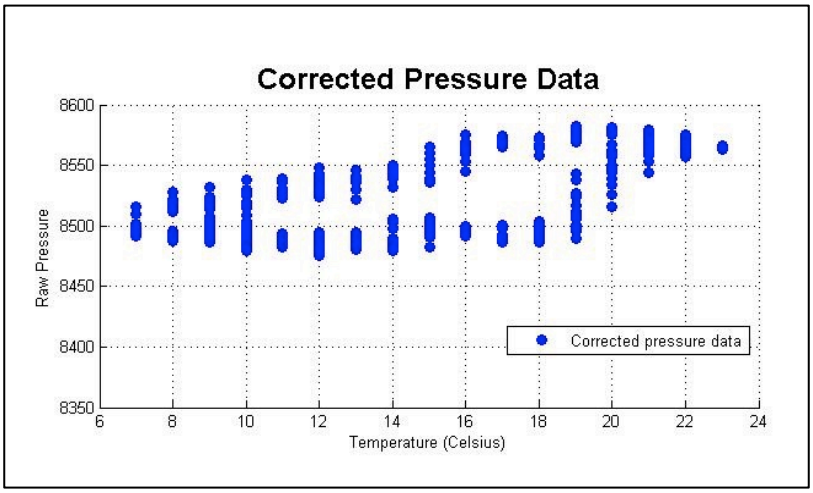



**Figure 14: Corrected pressure data when 2 liters of water were poured in the bottom storage of the LifeStraw.**

### 4.1.2 Detection of Events

The detection of water filter events is focused on detecting the regions of near constant slope in the raw data.

The steps taken to accomplish this task were:

- Re-sampled the non-uniformed distributed data in a uniformly manner.
- Detecting LifeStraw uses from the uniformly distributed data.

Re-sampled the data was necessary due to the missing data that gets lost during the transmission to the cloud and the battery discharges. To uniformly sample the data, Dr. Carson used a sliding window linear fit technique to calculate the slope of the raw data with a 20 minutes window of one-minute intervals [7].

The slope was calculated by creating a uniformly spaced time vector of one-minute intervals. Then, the timestamps that are within the current window are found making sure that there are at least seven points within the current window. This way, the slope $s[n]$

(Figure 15) and error $e[n]$ were estimated in a uniformly manner. The error was then normalized to be an indicator of the non-linearity of the slope data in each window.

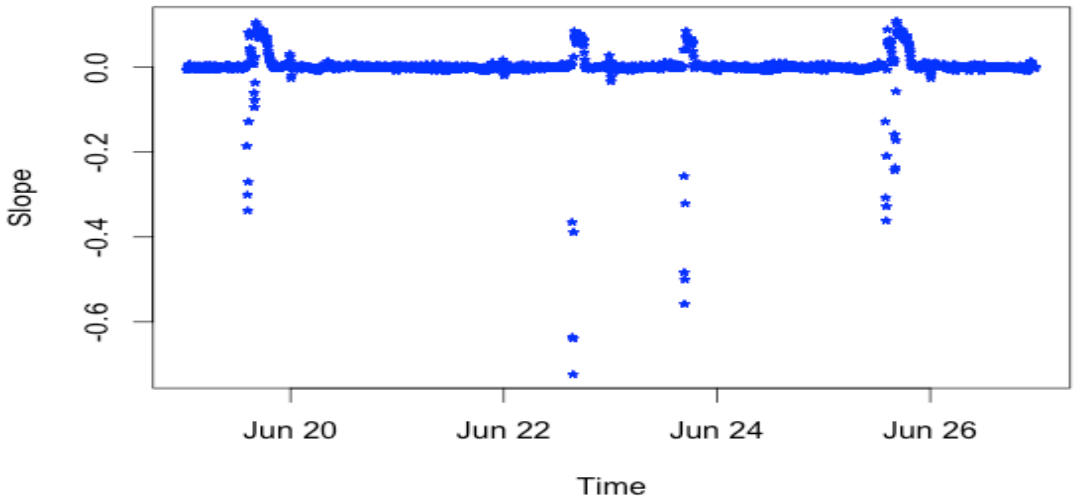



**Figure 15: Example of slope data distributed uniformly with a one min sample rate. The figure shows four clearly define events.**

After obtaining the error and slope, the LifeStraw uses were obtained by detecting regions with positive near-constant slope and low error. This was accomplished by calculating the slope spectrum of the raw data.

The slope spectrum is a tool that helps to visualize the slope over time. To construct the slope spectrum the range of slope that supposed to correspond to a filter usage is divided into a number of bins. A signal is created if the slope at a specified time is within that bin, i.e. slopes greater than 0.04. These signals are then penalized by the error $e[n]$ associated with the slope and convolved with a moving average to build up the spectrum [7].

$$s[m,n] = h_{MA} * (1 - e[n])$$

where, $e[n]$ is the error, and $h_{MA}$ the moving average for a length of 20 min (time of window), $n$ is the time index, and $m$ is the number of slope bins.

This process allows identifying uses of the LifeStraw as time interval by creating a binary signal, $x[n]$, that indicates which of the maximum of $s[m,n]$ values are higher than the threshold, 0.5.

$$x[n] = \begin{cases} 1 & \max_m S[m,n] \geq 0.5 \\ 0 & \text{otherwise} \end{cases}$$

The 1 and 0 signals represent the start and stop times of the filtering process. To identify when the interval $x[n]$ is high, Dr. Carson perform a first difference and then padded the intervals to an amount equal to the length of window multiplied by the threshold for accuracy in result.

***Improvements***

The improvements made to this part of the algorithm were primarily based on parameters and thresholds values used in the development of the algorithm. The improvements are:

- The size of the window: Originally the window size was of 20 minutes, but because pressure points are either missing or very high there are occasions in which there was not enough data to estimate the slope, which led to missing

events in the analysis. To solve this problem, the window size was changed to 30 minutes to get enough data to estimate the slope.

- Qualification for calculating the slope: Originally it was necessary to have at least 1/3 sample per minute, but as explained previously there were occasions in which there was not enough data to estimate the slope. Therefore it was changed to 1/10 sample per minute to facilitate slope calculation.
- Minimum and maximum values for slope ranges to be evaluated: The raw pressure data for the LifeStraw is observed to be noisy. This noise in the data interferes in the calculation of the slope and therefore it was necessary to define a range of slope values that represents a possible real event. Originally Dr. Carson chose slopes with values from 0.04 to 0.15. From visualization of multiple slopes figures, it was identified that for the algorithm to detect real events, the slope should be greater than 0.02.
- Threshold of $S[m,n]$: The threshold of $S[m,n]$ represents the minimum value for which the slope data is penalized by the error. Originally the threshold was 0.5 and was reduced to 0.4. This was to allow a little extra room to detect possible events.

### 4.1.3 Estimation of Water Volume

The amount of water filtered was estimated by calculating the flow rate of each specific event multiplied by the duration of it. The flow rate is then calculated by performing a linear fit on the pressure of each event to obtain the slope and then divide

this slope by the digital units per liter, in this case 225. The product of the event flow rate and its duration then estimates the volume of the event.

***Improvements***

As mentioned earlier one of the reasons behind the error in the analysis is the digital units per liter. Originally the value given by Dr. Carson was 225 digital units per liter. Laboratory experiments were conducted on three different LifeStraws to confirm or correct the value given by Dr. Carson. In the test five liters to one liter of water were filtered from each of the LifeStraws, and after the data was reported, the pressure readings were compared and an approximation of the pressure for each of the known volume values was obtained.

It was noticed that there was a difference in the pressure among the LifeStraws used for testing for same amount of water filtered. This was due to the fact that the sensors are not calibrated to a common standard i.e. the raw data of different LifeStraws presents different offsets. We use linear regression techniques to analyze this data. The regression analysis involves taking a set of data and fitting a trend of the data with the best-fit function. This allows determining the relation between the two variables. This is represented below:

$$Y(x) = ax + b,$$

where $Y(x)$ represents the difference in pressure of the initial and final value of the filtering event, $x$ represent the known volume added, the constants $a$ and $b$ represent the linear coefficient and y-intercept respectively.

Figure 16, shows the linear regression model of the data. The figure shows the function that fit the data and the R-square value, which gives an indication of how well the function fits the data [15]. This equation is shown below:

$$Y(x) = 200.58x - 77.28$$

$$R^2 = 0.93$$

From this analysis it was conclude that the digital pressure units per liter is 200.58 rather than 225 as suggested by Dr. Carson.

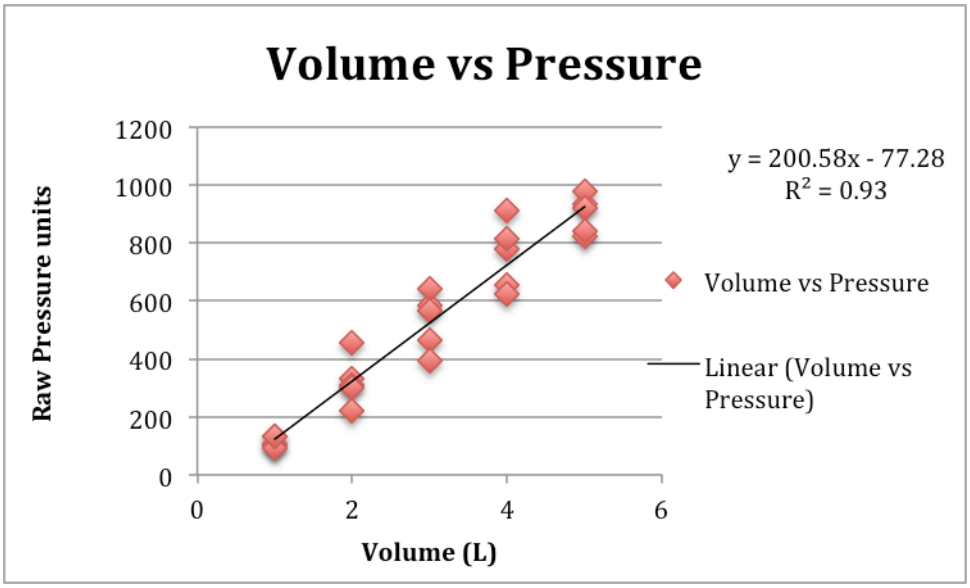



**Figure 16: Linear model of volume vs. units of pressure. There are 200.58 digital units of pressure per liter**

## 5. RESULTS AND DISCUSSION

This chapter provides an analysis of the results of the algorithm applied to laboratory data and field data. Section 5.1 presents the analysis of the results obtained from the baseline and updated algorithm using laboratory data. Section 5.2 presents the analysis of the first round of data collected in Rwanda using the updated algorithm, Section 5.3 gives a precision and recall analysis, and finally Section 5.4 provides the validation of the algorithm using a second round of data collected from Rwanda.

### 5.1 Analysis of Laboratory Data

After obtaining the number of events and estimated volume from the laboratory experiments, it was necessary to perform a regression analysis to identify how close the output fit into the expected result, and to check the improvement between the previous and the updated algorithm [9]. The following assumptions were taken into account in order to create a linear regression model:

1. The output volume from the algorithm relates to the known volume by a linear regression model

$$V_A = \beta_o + \beta_1 V_o + \epsilon$$

Where, $V_A$ is the output from the algorithm and $V_o$ is the known volume added.

2. The error, $\epsilon$, is independent and identically normally distributed with zero mean and constant variance, $N(0, \sigma^2)$.

Figure 17 and 18 present the scatter plots and regression models for the results with the updated and the previous algorithm. Each point in the diagrams represents an event. The purple line represents the fitted regression line and the blue lines the 95% prediction intervals. The prediction intervals represent the range where a single observation is likely to fall.

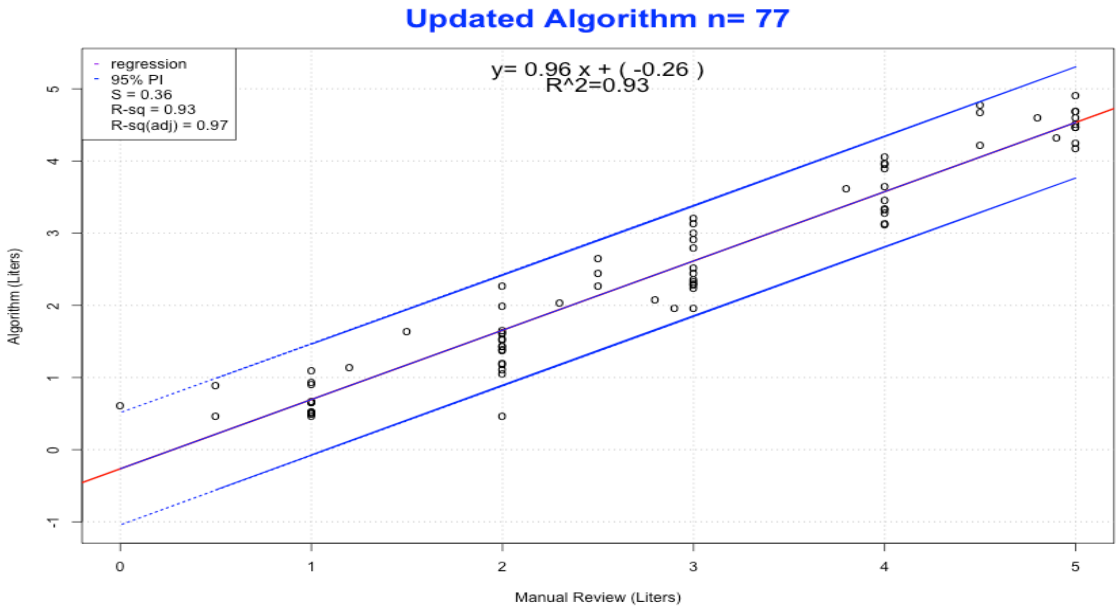



**Figure 17: Regression analysis of laboratory data using updated algorithm. Blue line represents the 95% prediction line. The figure shows that the output from the updated algorithm agrees more closely to the real value.**

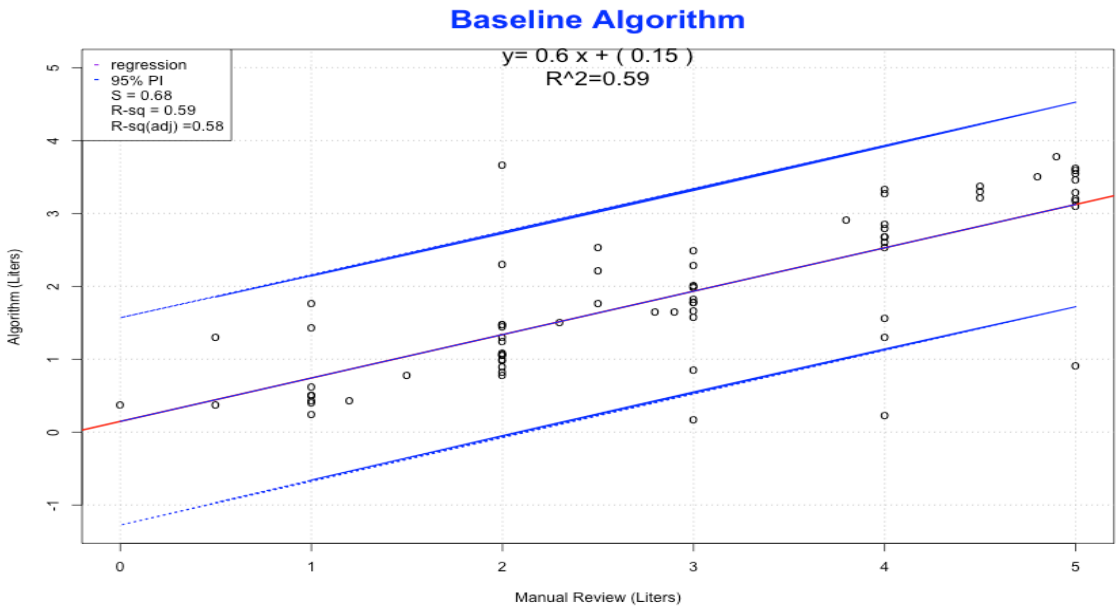



**Figure 18: Regression analysis of laboratory data using baseline algorithm. The figure shows that the output from the baseline algorithm is less accurate in estimating the real value. It presents higher variability with respect of the fitted line.**

Overall, it can be seen that the relationship between the known input volume and output for the algorithm is linear, but not perfect. From the updated algorithm, the average error across all the sensors under test is 16.2%, and from the baseline algorithm the total error is 39.6%. It is clearly visible that the updated algorithm agrees closely to the laboratory trials than the previous algorithm.

From the two regression models the main noticeable difference is the variability of the data around the regression line for the results of the baseline algorithm (Figure 17), which is much higher than Figure 16. The low $R^2$ and S (standard error of regression) gives a numerical explanation of this variability.

The regression model from the updated algorithm shows that for every liter of water filtered, it is expected that the output of the algorithm will increase by $0.96 \pm 0.03$ liters. The $R^2$ for the updated algorithm is 0.93, which indicates that the volume estimation is

more precise and concentrated within a range of 1.6 liters for an input of 1 liter. In comparison, the previous algorithm had a $R^2$ of 0.59 with a prediction interval that extended from -0.53 to 2.26 for an input of 1 liter, 2.79 liters.

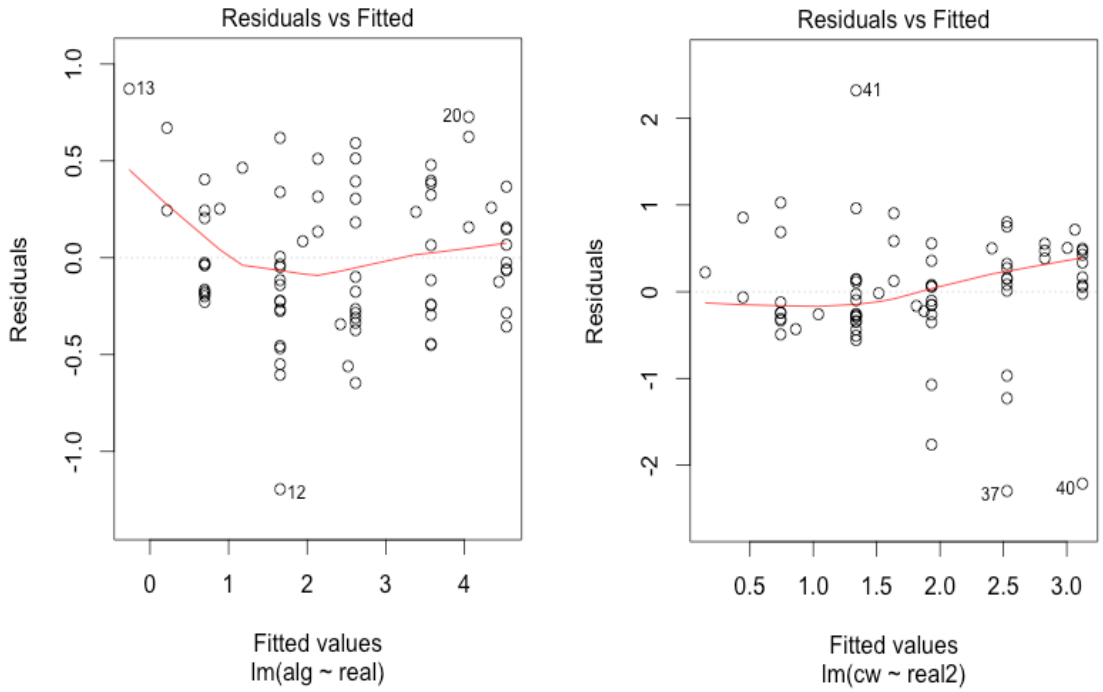



**Figure 19: Residual plots of updated and baseline algorithm. The figure in the left is from the updated algorithm and the figure in the right is from the baseline algorithm. The figure shows that the residuals from the updated algorithm are more symmetrical.**

Figure 19 shows the residual plots of the updated and baseline algorithm. The figure on the left represents the updated algorithm, and the figure on the right the baseline algorithm. The residuals, $\epsilon$, are the difference between the estimated volume value $y$ and the fitted values $\hat{y}$, and are defined as the error associated with that data. Each point in these plots is one event, where the x-axis represent the fitted value and the y-axis the residuals. The residual plots allow us to verify the correctness of the regression mode [10].

The residuals, $\epsilon$, are represented as:

$$\epsilon = y - \hat{y}$$

Positive values on the residuals mean that the estimated values are too low, and negative values mean that the estimated values are too high. For a good regression analysis the residuals should be symmetrical and concentrated around the mean.

The residual plots show that the residuals of the updated algorithm are more symmetrical and within values lower than one, as compared with the previous algorithm in which the residuals appear to be unbalanced to the left side.

### 5.2 Analysis of Field Data

A similar analysis was conducted with the field data to check the accuracy of the algorithm. To accomplish this task, manually approximations were used to identify the time and volume of the event. The analysis was conducted on 85 LifeStraw with approximate 933 candidate events considered.

In the manual review real events are expected to behave closely to what was observed in the laboratory. A detail explanation of the manual review process is shown in Appendix A.

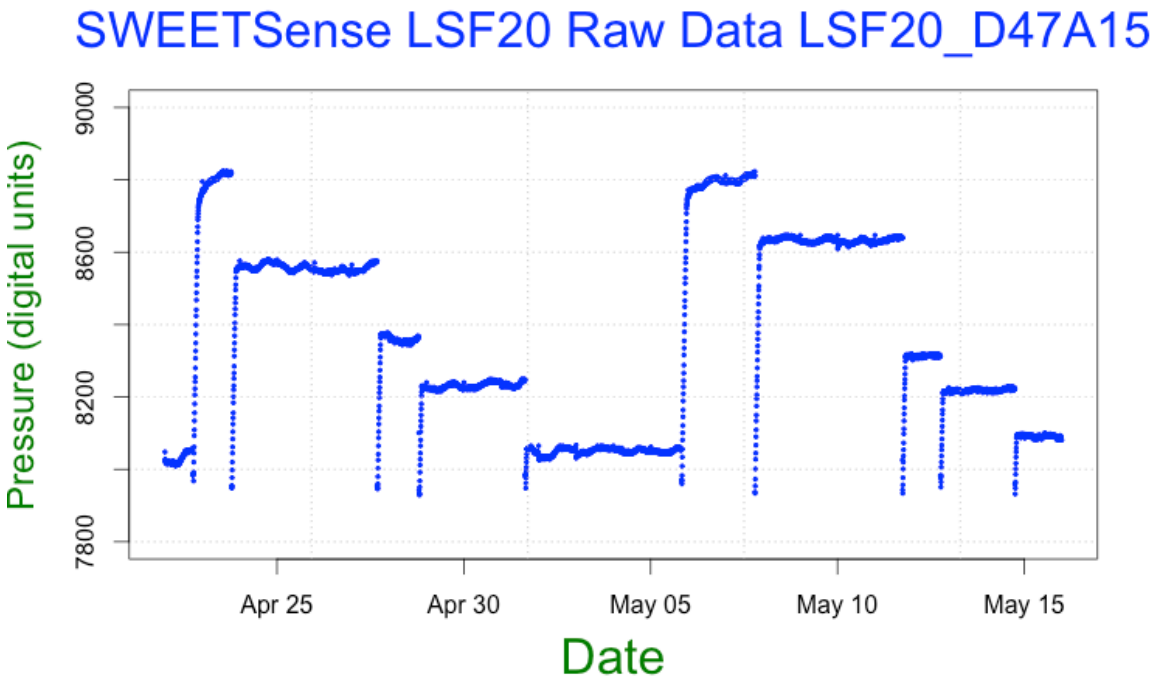



**Figure 20: Laboratory data. The figure shows data reported from a LifeStraw used in the Laboratory. The behavior of this data is used as a base to review the data from the field. In this figure there are clearly ten events well represented by a continuous increment of pressure over time. In the manual review we look for data that simulates this behavior.**

Figure 21 shows the regression analysis of field data analyzed with the updated algorithm. This plot shows the relationship between the output volume of the algorithm and the manual review of the field data. Each point in the figure corresponds to an event and the blue lines on the sides represents the 95% prediction lines.

Overall it can be seen that the algorithm agrees closely with the observed events. The total average error across all LifeStraw data from Rwanda is 13.2%.

The regression model shows that for every liter of water filtered the output of the algorithm will increase by $1.039 \pm 0.005$ liters. The $R^2$ for the updated algorithm is 0.97 with a prediction interval ranging from 0.54 to 1.50 for an input of one liter of water, i.e.

0.96 liters. These indicate the low variability and accuracy of the algorithm when estimating volume.

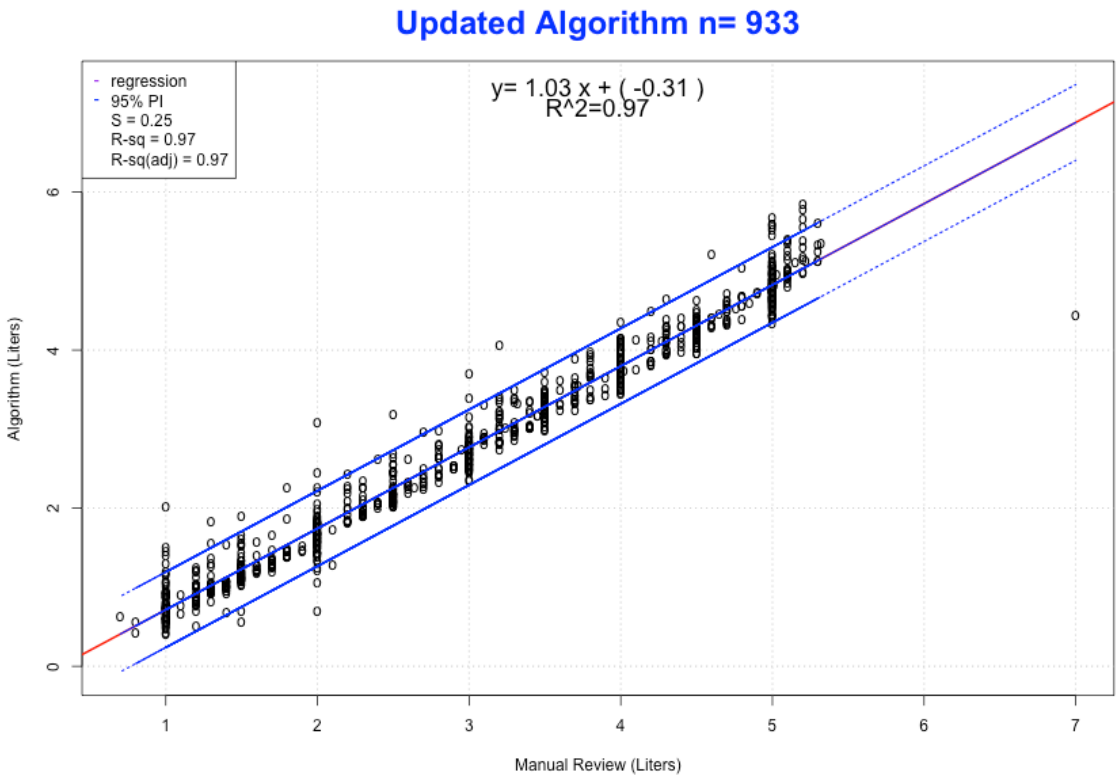



**Figure 21: Regression analysis of field data using updated algorithm. Y-axis indicates the algorithm output and the X-axis indicates the manual review. The figure shows that from 933 observations the output from the algorithm closely agrees with the manual review.**

While the results from the improved algorithm significantly reduced observed error, observation of the LifeStraw and the overall results of the algorithm suggested a minimum detection limit of approximately 0.5 liters. Therefore, to solve this problem 0.5 liters were added to the output volume. By doing that, it was noticed that the results were closer to the manual review and the error was reduced to 10.5%. The $R^2$ is still 0.97 and the prediction interval went down to 0.9 liters.

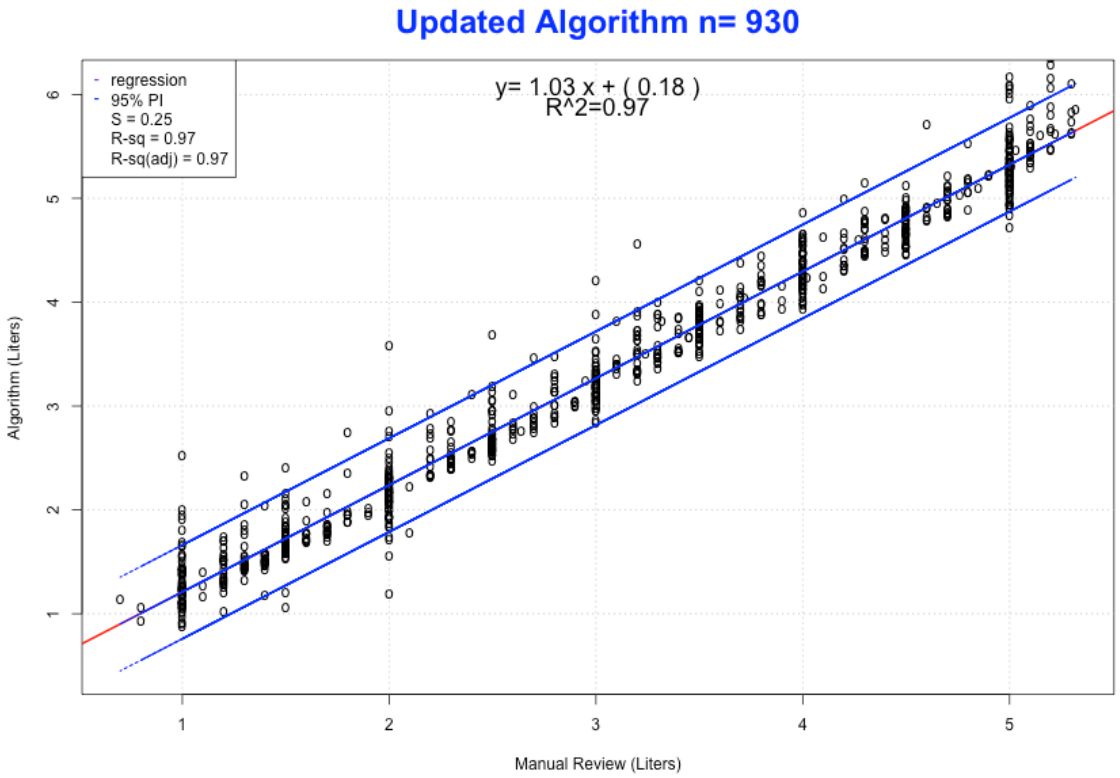



**Figure 22: Regression analysis of field data after improvement in updated algorithm. There was an improvement in the results after adding 0.5 liters to the estimated volume of water**

Overall, it is clearly shown that the detection and estimation of volume from the updated algorithm agrees closely with the real values whereas the results from the previous algorithm show more variation. It was observed from the regression analysis that the updated algorithm has low variability in the estimation of the volume, which means that the results are more precise. From the laboratory results the average error for the data is 11.5% as compared with the baseline algorithm, which is noticeably higher, 39%. The field data presented an average error of 10.5%. Errors may be attributed to a lack of calibration of each individual sensor, electronic noise, ambient temperature, and variations in real-world use of the water filters compared to laboratory approximations.

### 5.3 Precision and Recall

Precision and recall are terms generally used to evaluate the performance of a model [15]. In this thesis precision and recall are used to evaluate the accuracy of the algorithm in identifying events. Precision or positive predictive value, refers as a measure of the exactness of the algorithm in identifying events:

$$Precision = \frac{TP}{TP + FP}$$

Recall, also known as sensitivity, refers to the ability of the test to correctly identify true-events:

$$Recall = \frac{TP}{TP + FN}$$

where,

TP (True Positive): The algorithm accurately detects a real event.

FP (False Positive): The algorithm detects an event, but in reality there was not real event.

TN (True Negative): The algorithm doesn't detect an event because there was no event.

FN (False Negative): The algorithm doesn't detect an event, but in reality there was an event.

The results obtained with the algorithm and manual review are presented in the confusion matrix shown in Table 3. Each cell in the confusion matrix contains the number of incorrect or correct results obtained with the algorithm. The first column in the table refers to the numbers of true positives and false negatives, and the second column refers to the number of false positives and true negatives. The number of true negatives is not found on this application due to the difficulty in estimating the number of times in which the LifeStraw was not in used.

Table 3 shows the results from the algorithm for 85 LifeStraws deployed in Rwanda. The table shows that 979 events were detected with the algorithm, from which 930 are true positives and 49 and false positives. It also shows 10 false negatives. The precision of the algorithm is 930/(930+49) or 0.95, the recall or sensitivity is 930/(930+10) or 0.98.

**Table 3: Algorithm Confusion Matrix (Round 1)**

| | Test Results | | |
|---|---|---|---|
| | **True** | **False** | **Total** |
| **Event** | 930 | 49 | 979 |
| **No Event** | 10 | -- | 10 |
| **Total** | 940 | 49 | 989 |

## 5.4 Validation

This section discusses the validation of the algorithm. The validation of the algorithm tries to investigate the accuracy of the algorithm after all the necessary improvements. To accomplish this task, the algorithm was applied to data collected from 80 LifeStraws and compared with a manual review of this data. The 80 LifeStraws are the second round of LifeStraws deployed in Rwanda; the first round was analyzed in section

5.2. The manual review of these datasets was done in a similar way as section 5.2. This second round of data was not initially considered in the algorithm development stage.

Figure 23 presents the regression analysis of the field data used in round 2. The figure shows that for every liter of water filtered the output of the algorithm increases by $1.036 \pm 0.009$ liters. The $R^2$ is 0.97 and the prediction interval for an input of two liters is 1.05 liters. The total average error was 5%. As with round 1 analysis (section 5.2), the output from the algorithm has significantly improved over the baseline.

Table 4 presents the total number of events detected with the algorithm and the manual review. The table shows that there are 570 events detected with the algorithm of which 519 are true positives and 51 are false positives. The number of true negatives was not found for the same reason as explained in section 5.2.

Based on the results presented in the Table 4, the sensitivity or recall of the algorithm is $\frac{519}{528} = 0.98$ and the precision is $\frac{519}{570} = 0.91$.

**Table 4: Algorithm Confusion Matrix (Round 2)**

| | **Test Results** | | |
|---|---|---|---|
| | **True** | **False** | **Total** |
| **Event** | 519 | 51 | 570 |
| **No Event** | 9 | -- | 9 |
| | 528 | 51 | 579 |

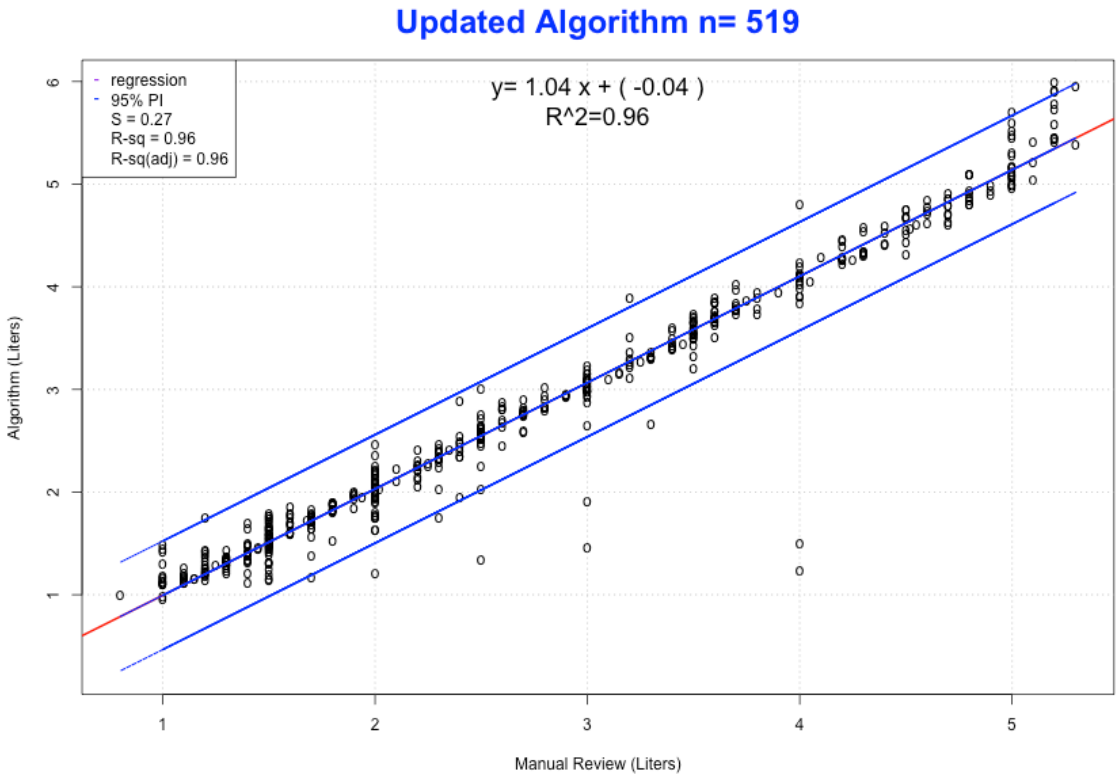



**Figure 23: Regression analysis of field data for round 2. The figure shows that 519 true events were detected with the algorithm with an average error of 5%. The $R^2$ is 0.97 and the prediction interval for an input of two liters is 1.05 liters.**

Overall, the results obtained from the algorithm are fairly accurate and seemed to closely correlate with the expected values. 99% of the events were detected with 5% of total average error. Figure 24 shows an example of the analysis of water pressure data. The figure highlights the events detected by the algorithm and the amount of water estimated.

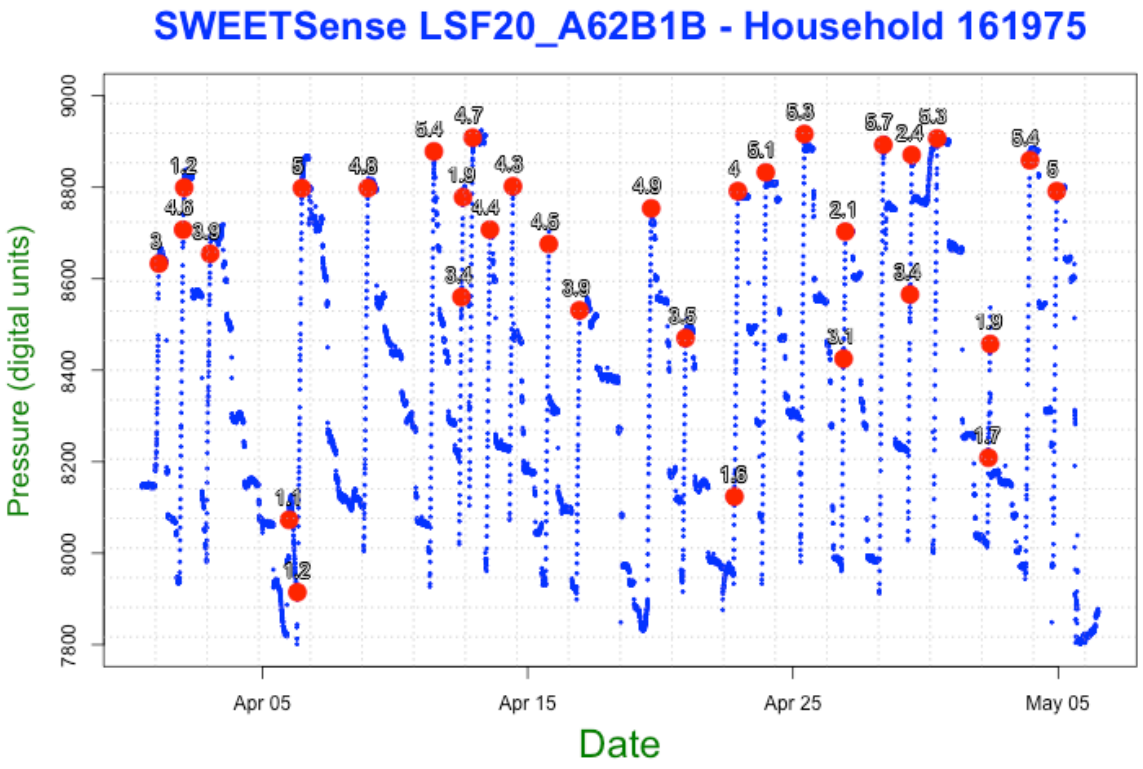



**Figure 24: Example of water pressure data from sensor installed in LifeStraw water filter safe water storage container (blue scatter points are digital pressure units); detected events are represented by the red circle and the estimated volume is highlighted by the numbers on top.**

# 6. SUMMARY AND CONCLUSION

This thesis discusses the refinement and validation of algorithms for data collected from pressure sensors that are used in water filters. The objective behind this thesis is to accurately detect and estimate the volume of water of filling events, also validate the analysis using laboratory data, and lastly analyze field data and validate algorithm.

The success of this project will allow demonstrating in an accurate manner the effectiveness of using the Sweet lab's cellular reporting instrumentation system to monitor the behavior of household water filters.

## 6.1 Summary

The need to acquire objective data for development programs in developing countries has brought the attention to the use of sensors as a tool to overcome the challenges of household-to-household surveys. Sensors have demonstrated to be beneficial to monitor and evaluate the successful rate of programs that measure behavior and functionality of these types of programs.

The SweetLab at Portland State University developed a cellular reporting instrumentation system to provide objective data in programs such as, water filters, water-carrying backpacks, cook stoves, and many other devices. This thesis primarily discussed the analysis of data collected with pressure transducer used in household water filters that are deployed in Rwanda.

The analysis is based upon an algorithm developed by Dr. Carson Wick, this algorithm involved three parts: preprocessing raw data, detection of events, and estimation of volume. Each of these parts was slightly modify to allow better results

while detecting filtration events, and estimating volume. The modifications on the algorithm were based on the hypothesis that the error in the analysis could be due to wrong estimation of slope of the event, start and stop time of the event, or the digital units per liter, i.e. the expected change of the pressure reading per liter.

As discussed in Chapter 3 Laboratory experiments were conducted to collect data and assist in the refinement of the algorithm, field data was also used to make corrections as necessary and validate the algorithm.

The analysis of data and improvements made to the algorithm are discussed in Chapter 4. The modifications on the algorithm were related to: the compensation of the temperature in which collection data and regression techniques were used to allow reducing the effect of the temperature on the pressure data. The calculation of the slope was improved by modifying parameters and thresholds that allow easier and better estimation of the slope. Also, we were able to obtain a better the estimation of volume by redefining the relationship between the digital units of pressure and liters of water, and by adding 0.5 liters to the final result of the estimation.

In chapter 5 we presented a comparison of results of the improved and baseline algorithm, a validation of the algorithm using a second round of datasets of LifeStraws deployed in Rwanda, and a precision and recall analysis to evaluate the accuracy of the algorithm.

## 6.2 Conclusion

The results presented in this thesis demonstrate that the improvements made to the algorithm are more accurate as compared to the baseline algorithm. Overall there was a greater improvement detecting real events and the estimation of volume was more precise. Originally the error presented in the algorithm was 39% for laboratory analysis, which came down to 11.5% after the necessary modification. Analysis of field data presented an error of 5%. The manual review in the latest stages of the work was more precise due to the better understanding of the data, this also contributed to the low error in the field data analysis.

Errors in the analysis may be attributed to real-world behavior of the water filter, electronic noise, ambient temperature, and variations in the approximation made to the field data.

For further studies and improvement of the results it is important to try to simulate this behavior to try to get a better understanding of the performance of the data. Additionally, to be able to accurately correlate the results of the algorithm with the expected value it is important to know the real value of the water filtered.

## 7. REFERENCES


1. UNICEF. *Child survival fact sheet: Water and sanitation*; UNICEF Fact sheet; UNICEF: New York, NY, USA, 2004
2. *LifeStraw Family 2.0 by Vestergaard.* Retrieved from: <http://www.vestergaard.com/index.php/ourproducts/lifestraw/item/lifestraw-family-2-0/>.
3. Barstow CK, Ngabo F, Rosa G, Majorin F, Boisson S, et al. (2014) Designing and Piloting a Program to Provide Water Filters and Improved Cookstoves in Rwanda. PLoS ONE 9(3): e92403.doi:10.1371/journal.pone.0092403
4. PackH2O. Retrieved from: http://www.packh2o.com/learn
5. Agua Para Vivir." *PackH2O*. Web. <http://www.packh2o.com/gallery/docs/>.
6. Portland State University Sweetlab. Available online: http://www.pdx.edu/sweetlab/delagua-rwanda
7. Thomas, E; Zumr, Z; Graf, J; Wick, C; McCellan, J; Iman, Z; Barstow, C; Spiller, K; Fleming, M. Remotely Accessible Instrumented Monitoring of Global Development Programs: Technology Development and Validation. *Sustainability* 2013, 5, 3288-3301
8. Roumis, D; Rostapshova, O. "Sensors for MERL: What Works? What Does Not? What Have We Learned?" ICTworks, 23 Nov. 2015. Web. 28 Dec. 2015.
9. Ferrari, D; Head, T. Regression in R. UCLA Department of Statistics Consulting Center. Feb 10, 2010
10. "Interpreting Residual Plots to Improve Your Regression." *Interpreting Residual Plots to Improve Your Regression*. N.p., n.d. Web. 11 sep. 2015.
11. Weather in Rwanda. Retrieved from: http://www.accuweather.com/en/rw/kigali/293211
12. Institute for Statistics and Mathematics. Available online: http://www.r-project.org
13. Palmer, J, Precise Pressure Sensor Temperature Compensation Algorithms (2006). *Thesis. Binghamton University.*
14. Recktenwald, Gerald, "Least Squares Fitting of Data to a Curve", Portland State University, November 2001.
15. "Classification Accuracy Is Not Enough: More Performance Measures You Can Use." Machinelearningmastery. Jason Brownlee. Web. 8 May 2016.
16. Mercado, I. Temperature dataloggers as stove use monitors (SUMs): Field methods and signal analysis. *Biomass Bioenergy* 2012, *47*, 459–468.
17. Watras, C.J., P.C. Hanson, T.L. Stacy, K.M. Morrison, J. Mather, Y-H Hu and P. Milewski. 2011. A temperature compensation method for CDOM fluorescence sensors in freshwater. Limnol. Oceanogr: Methods 9: 296-301.

## APPENDIX A: MANUAL REVIEW FIELD DATA

The manual review consisted of a detail observation of the data sets collected in the field to identify a continuous diagonal line, which describes the increment of pressure over time, and a roughly approximation of the volume of the water.

The estimation of the volume of water in the manual review was done as follow:

1. Collected multiple pressures readings for multiple known volumes values in the laboratory.

| Volume (L) | Final Pressure | Initial Pressure | Time |
|---|---|---|---|
| 1 | $P_{2a}$ | $P_{1a}$ | $T_{1a}$ |
| 2 | $P_{2b}$ | $P_{1b}$ | $T_{1b}$ |
| 3 | $P_{2c}$ | $P_{1c}$ | $T_{1c}$ |
| 4 | $P_{2d}$ | $P_{1d}$ | $T_{1d}$ |
| : | : | : | : |
| : | $P_{2n}$ | $P_{1n}$ | $T_{1n}$ |

2. Obtained the pressure values that represent the start and stop values of the water filter use event. These values can be obtained from the algorithm.
3. Estimated the delta value of each of these filtering events by subtracting the pressure values that represent the start and stop times.

$$\Delta P = Final\ Pressure - Initial\ Pressure$$

4. Plotted the delta values of the pressure readings against the known liters of water.
5. Fitted a line through the data to obtain a mathematical regression that relates the delta pressure value with the volume.
6. Used the following equation to get a roughly approximation of the volume.

$$V = \frac{77.28 + \Delta P}{200.58}$$

7. Observed the data to verify the results and modified them if needed. For the observation, it was necessary to verify the initial and final pressure value for each event to make sure the algorithm accounted for the whole length of the event. Also check the correlation between adjacent events.

The figures below (Figure 25 and 26) present examples of the manual review for LifeStraws deployed in Rwanda. The examples show results from the baseline algorithms to demonstrate the need to improve the algorithm to obtain more accurate results.

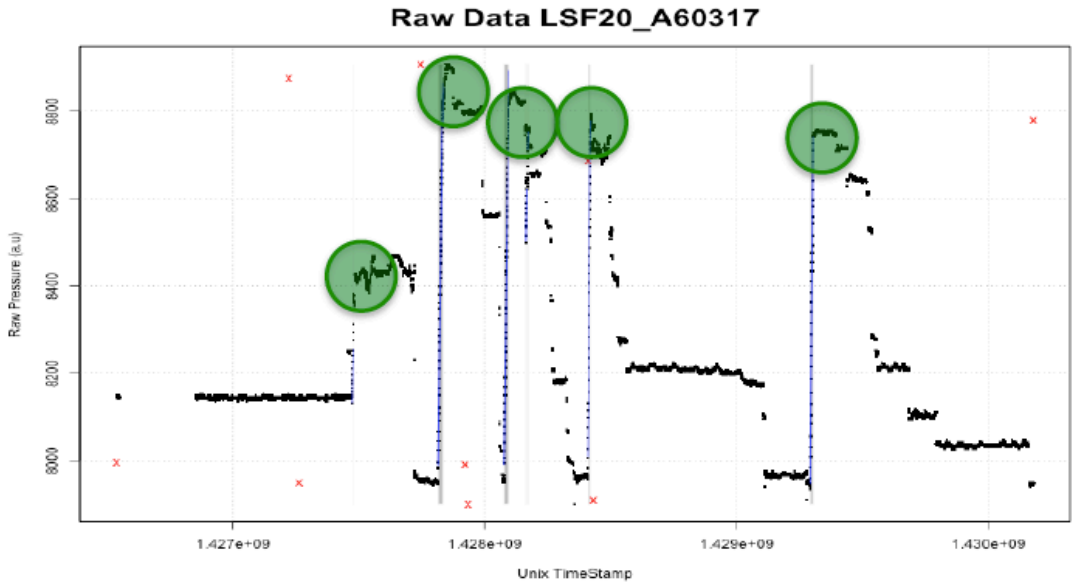



**Figure 25: Analysis of field data (LSF20_A60317) with baseline algorithm. The analysis shows that there are nine events, but in reality there are five. A green circle highlights the manual review.**

The manual review for LifeStraw unit LSF20_A60317 suggests that there are five real events instead of nine events as implied by the baseline algorithm. The estimation of the volume of water was done by using the equation presented in part 6 and by modifying those results accordingly to the observation of the data. The detail analysis of the manual review is shown below:

1. First event:
    - Use equation to approximate volume:

$$V = \frac{77.28 + 125.8}{200.58} = 1.01$$

    - Review results by observing data: From the figure it was noticeable that the real event is longer and the final pressure value is higher than what was predicted by the algorithm, approximately 8400. Therefore the final result of the volume was:

$$\Delta P = 8400 - 8130.91 = 269.1$$

$$V = \frac{77.28 + 269.1}{200.58} \sim 1.7$$

2. Second event:
    - Use equation to approximate volume:

$$V = \frac{77.28 + 842.08}{200.58} = 4.5$$

    - Review results by observing data: From the figure one can notice that the final pressure of the event is a little bit higher than what was predicted by the algorithm, but it is difficult to estimate the value by observation. Therefore if

we checked the correlation with the next event (third event), with volume 4.7, one can assume that the value for the second event is higher than 4.7

3. Third event:
    - Use equation to approximate volume:

$$V = \frac{77.28 + 876.06}{200.58} \sim 4.75$$

4. Fourth event:
    - Use equation to approximate volume:

$$V = \frac{77.28 + 715.04}{200.58} = 3.95$$

    - Review results by observing data: The figure suggested that the final pressure value is higher than what was predicted by the algorithm, approximately 8800. Therefore the final result of the volume was:

$$\Delta P = 8800 - 8008.84 = 791.16$$

$$V = \frac{77.28 + 791.16}{200.58} \sim 4.32$$

5. Fifth event:
    - Use equation to approximate volume:

$$V = \frac{77.28 + 794.37}{200.58} = 4.35$$

Table 5 summarizes the results of the manual review for the LifeStraw unit LSF20_A60317. The table shows the volume estimated by the baseline algorithm, the

final and initial pressure of each of the events detected, the delta value of the pressure, and the volume estimated by the manual review.

**Table 5: Manual Review LSF20_A60317**

| Baseline Volume (L) | Final Pressure (a.u) | Initial Pressure (a.u) | Δ Pressure (a.u) | Manual Review Volume (L) |
|---|---|---|---|---|
| 0.56 | 8256.71 | 8130.91 | 125.80 | 1.7 |
| 3.89 | 8864.56 | 8022.48 | 842.08 | 4.8 |
| 0.37 | 8854.43 | 8769.89 | 84.55 | NA |
| 4.02 | 8893.76 | 8017.70 | 877.06 | 4.7 |
| 0.53 | 8621.78 | 8501.47 | 120.32 | NA |
| 0.19 | 8754.31 | 8710.67 | 43.64 | NA |
| 3.17 | 8723.88 | 8008.84 | 715.05 | 4.3 |
| 0.23 | 8778.13 | 8726.07 | 52.06 | NA |
| 3.53 | 8740.97 | 7946.61 | 794.37 | 4.3 |

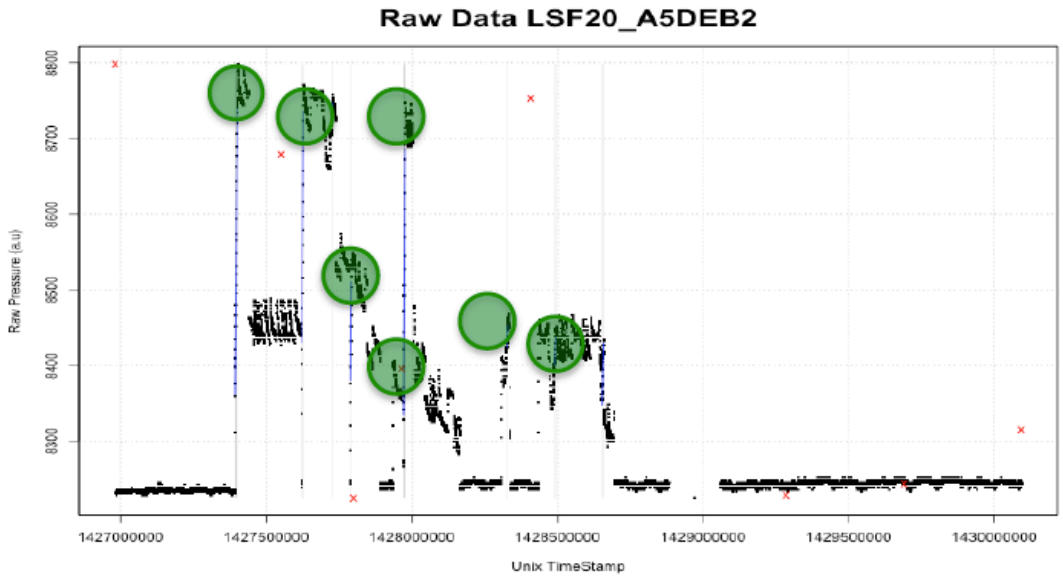



**Figure 26: Analysis of field data (LSF20_A5DEB2) with baseline algorithm. The green circles highlight the manual review.**

The manual review for LifeStraw unit LSF20_A5DEB2 suggests that there are seven real events instead of eight events as implied by the baseline algorithm. The estimation of

the volume of water was done as shown in the previous example and the summary of the results is shown in Table 6.

**Table 6: Manual Review LSF20_A5DEB2**

| Baseline Volume (L) | Final Pressure (a.u) | Initial Pressure (a.u) | Δ Pressure (a.u) | Manual Review Volume (L) |
|---|---|---|---|---|
| 1.69 | 8740.12 | 8226.03 | 514.09 | 3 |
| 1.36 | 8735.76 | 8251.46 | 484.3 | 2.9 |
| 0.12 | 8746.32 | 8719.32 | 27 | NA |
| 0.59 | 8510.96 | 8221.08 | 289.88 | 1.8 |
| 1.56 | 8686.40 | 8273.32 | 413.08 | 2.8 |
| 0.13 | 8453.39 | 8246.20 | 207.19 | 1.4 |
| 0.17 | 8437.18 | 8209.18 | 228 | 1.5 |
| 0.37 | 8429.47 | 8347.08 | 82.39 | NA |

## APPENDIX B: LABORATORY TEST PLAN

### Introduction

The test plan documents and tracks the necessary information required to effectively define the approach to be used in the testing of the LifeStraw algorithm.

This test plan is designed to test and calibrate the R algorithm with regards to:

- Performance
- Functionality
- Reliability

The objective of this algorithm is to detect the daily uses of the LifeStraw and to estimate the amount of water filtered by the user.

### Test Schematic

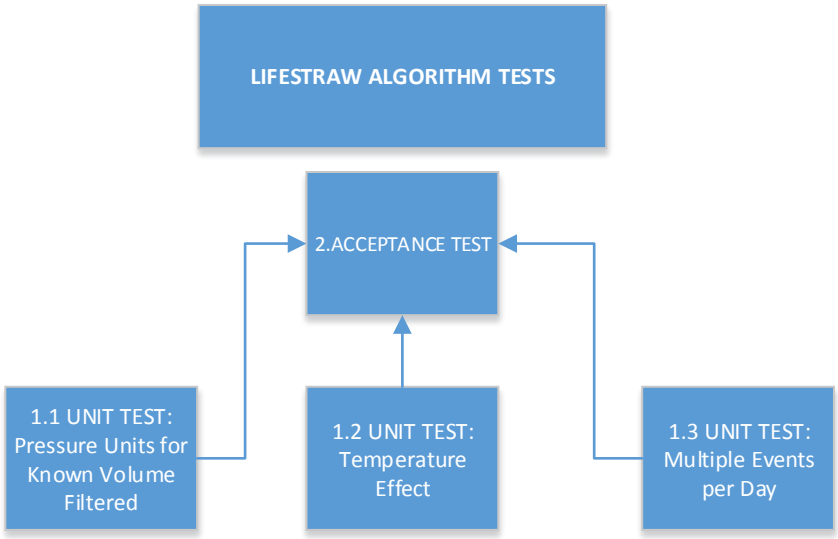



**Figure 27: LifeStraw algorithm tests schematic**

**Test Cases Overview**

1. **Unit Test**

**1.1 Pressure Units for known Volume**

Determine pressure units of known volume of water filtered and compare with other LifeStraw units.

**1.2 Temperature Effect in Raw Pressure Data**

Investigate the temperature effect in the pressure readings.

**1.3 Multiple Events per Day**

Check if the algorithm is able to detect events that happen more than once a day.

2. **Acceptance test**

Final test after all the necessary modifications are added to the algorithm. The acceptance test verifies that the algorithm meets the requirements.

## Test Cases

Unit Test Cases

- Pressure Units for Known Volume

| **Test Case Name** | Unit Test – Pressure Units for Known Volume | | | | **Test ID#:** | LSF20 -01 |
|---|---|---|---|---|---|---|
| **Description:** | Determine the pressure units for known volumes of water filtered and check whether or not the pressure readings of the LifeStraws under test is approximately the same when the same amount of water is filtered | | | | | |
| **Setup:** | 1. Pour five liters of water on the upper storage tank of different LifeStraws. Record time.<br>2. Drain the water from bottom storage tank the following day.<br>3. Repeat steps 1 and 2 for four, three, two, and one liters of water. | | | | | |
| **LifeStraw ID** | **Volume In** | **Time In** | **Volume Out** | **Time Out** | **Pressure** | **Comments** |
| LSF20_D47A 15 | 5 | 4/13/15 17:55 | 4.5 | 4/14/15 19:55 | 9124.36 | |
| | 4 | 4/14/15 20:07 | 4 | 4/16/15 15:05 | 8675.46 | |
| | 3 | 4/16/15 15:12 | 3 | 4/20/15 17:29 | 8538.44 | |
| | 2 | 4/20/15 17:38 | 2 | 4/21/15 16:53 | 8280.29 | |
| | 1 | 4/21/15 17:02 | 1 | 4/22/15 17:47 | 8063.86 | |
| | 5 | 4/22/15 18:10 | 5 | 4/23/15 18:39 | 8791.16 | |
| | 4 | 4/23/15 18:53 | 4 | 4/27/15 16:16 | 8588.70 | |
| | 3 | 4/27/15 16:30 | 3 | 4/28/15 19:09 | 8408.28 | |
| | 2 | 4/28/15 19:26 | 2 | 5/1/15 15:22 | 8246.58 | |
| | 1 | 5/1/15 15:33 | 1 | 5/5/15 19:44 | 8074.42 | |
| | 5 | 5/5/15 19:50 | 4.1 | 5/7/15 18:55 | 8786.51 | |
| | 4 | 5/7/15 18:58 | 3.4 | 5/11/15 17:40 | 8674.14 | |
| | 3 | 5/11/15 17:47 | 2.5 | 5/12/15 18:08 | 8364.14 | |
| | 2 | 5/12/15 18:11 | 1.3 | 5/14/15 17:55 | 8237.41 | |
| | 1 | 5/14/15 18:00 | 0.5 | 5/22/15 14:05 | 8122.99 | |

| | | | | | | |
|---|---|---|---|---|---|---|
| LSF20_D47D0D | 5 | 4/13/15 17:56 | 5 | 4/14/15 19:59 | 9130.81 | |
| | 4.5 | 4/14/15 19:57 | 4.2 | 4/16/15 15:07 | 9071.21 | |
| | 3 | 4/16/15 15:11 | 3 | 4/20/15 17:33 | 8798.11 | |
| | 2 | 4/20/15 17:40 | 2 | 4/21/15 16:55 | 8655.41 | |
| | 1 | 4/21/15 16:57 | 1 | 4/21/14 17:49 | 8307.67 | |
| | 5 | 4/22/15 18:08 | 5 | 4/23/15 18:37 | 9147.92 | |
| | 4 | 4/27/15 16:28 | 4 | 4/28/15 19:01 | 9066.78 | |
| | 3 | 4/28/15 19:20 | 3 | 5/1/15 15:21 | 8708.72 | |
| | 2 | 5/1/15 15:32 | 2 | 5/5/15 19:47 | 8462.01 | |
| | 1 | 5/5/15 19:49 | 1 | 5/7/15 18:55 | 8300.19 | |
| | 4.8 | 5/7/15 18:57 | 4.8 | 5/11/15 17:43 | 9095.23 | |
| | 4 | 5/11/15 17:46 | 4 | 5/12/15 18:05 | 8912.29 | |
| | 3 | 5/12/15 18:10 | 3 | 5/14/15 17:57 | 8644.45 | |
| | 2 | 5/14/15 17:59 | 2 | 5/18/15 16:53 | 8451.18 | |
| | 1 | 5/18/15 16:56 | 1 | 5/22/15 14:30 | 8250.00 | |

| | | | | | | |
|---|---|---|---|---|---|---|
| LSF20_A6340A | 4.5 | 4/13/15 17:54 | 4.5 | 4/14/15 19:51 | 8979.17 | |
| | 4 | 4/14/15 20:12 | 4 | 4/16/15 15:02 | 8768.17 | |
| | 3 | 4/16/15 15:13 | 3 | 4/20/15 17:26 | 8602.96 | |
| | 2 | 4/20/15 5:37 | 2 | 4/21/15 16:50 | 8410.64 | |
| | 1 | 4/21/15 17:00 | 1 | 4/22/15 17:45 | 8277.53 | |
| | 4.9 | 4/22/15 18:16 | 4.9 | 4/23/15 18:47 | 9001.58 | |
| | 4 | 4/23/15 18:57 | 4 | 4/27/15 16:18 | 8930.59 | |
| | 2.9 | 4/27/15 16:31 | 2.9 | 4/28/15 19:13 | 8572.26 | |
| | 2 | 4/28/15 19:27 | 2 | 5/1/15 15:27 | 8392.52 | |
| | 1 | 5/1/15 15:33 | 1 | 5/7/15 18:50 | 8253.35 | |
| Overall test result | | | | | | |

- Temperature Effect in raw data

| **Test Case Name** | Unit Test – Temperature effect in raw data | | | **Test ID#:** | LSF20 -02 |
|---|---|---|---|---|---|
| **Description:** | Investigate the effect of temperature in the pressure readings. | | | | |
| **Setup:** | 1. Pour five litters of water in one LifeStraw placed outside. Record time<br>2. Drain the water from the bottom storage tank after two days. Record time.<br>3. Repeat steps 1 and 2 for four, three, two, and one liters. | | | | |
| **LifeStraw ID** | **Volume In** | **Time In** | **Volume Out** | **Time Out** | **Pressure** | **Comments** |
| LSF20_A6340A | 5 | 5/8/15 9:39 | 5 | 5/12/15 22:13 | 9074.37 | |
| | 4 | 5/12/15 22:16 | 4 | 5/14/15 22:14 | 8882.93 | |
| | 3 | 5/14/15 21:35 | 2.5 | 5/20/15 23:55 | 8568.54 | |
| | 3 | 5/20/15 0:00 | 3 | 5/22/15 22:00 | 8624.00 | |
| | 2 | 5/22/15 22:05 | 2 | 5/24/15 23:30 | 8260.43 | |

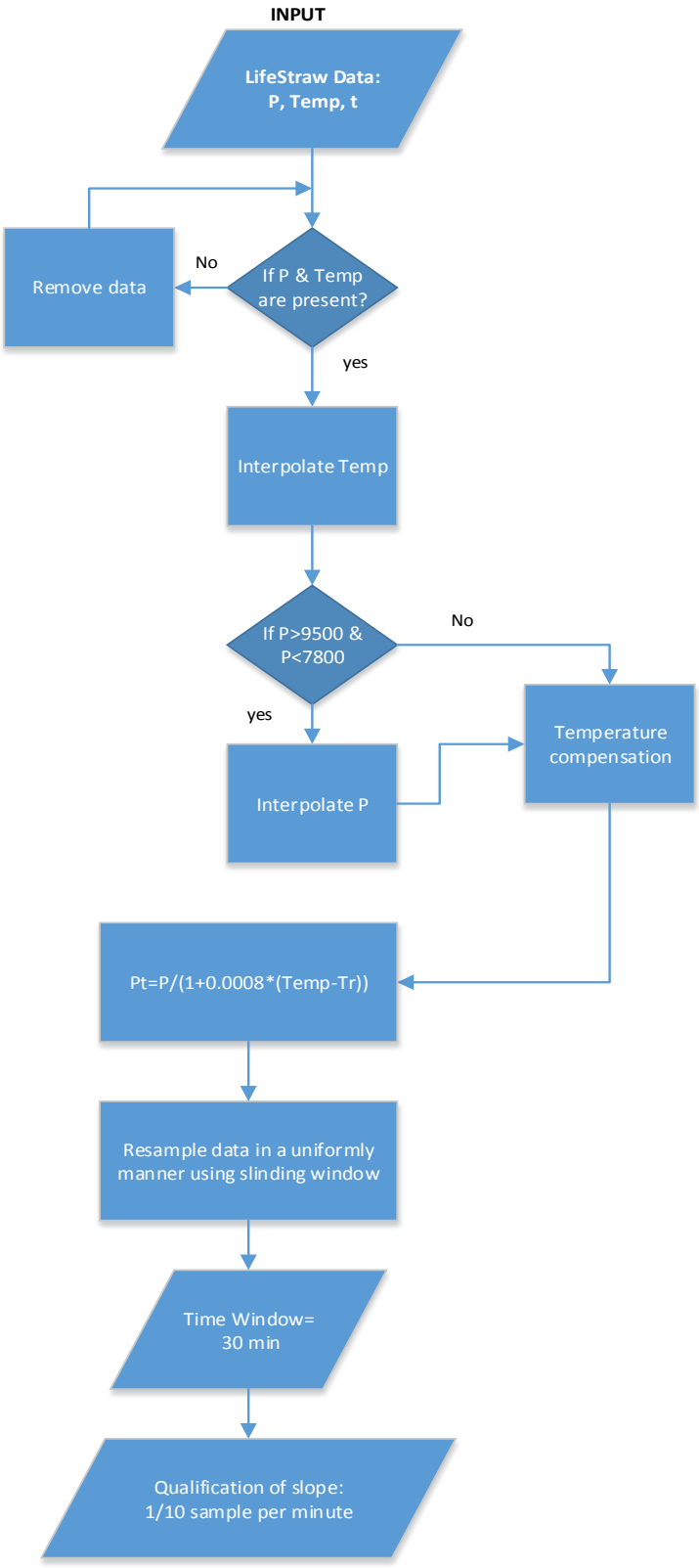
- Multiple Events per Day

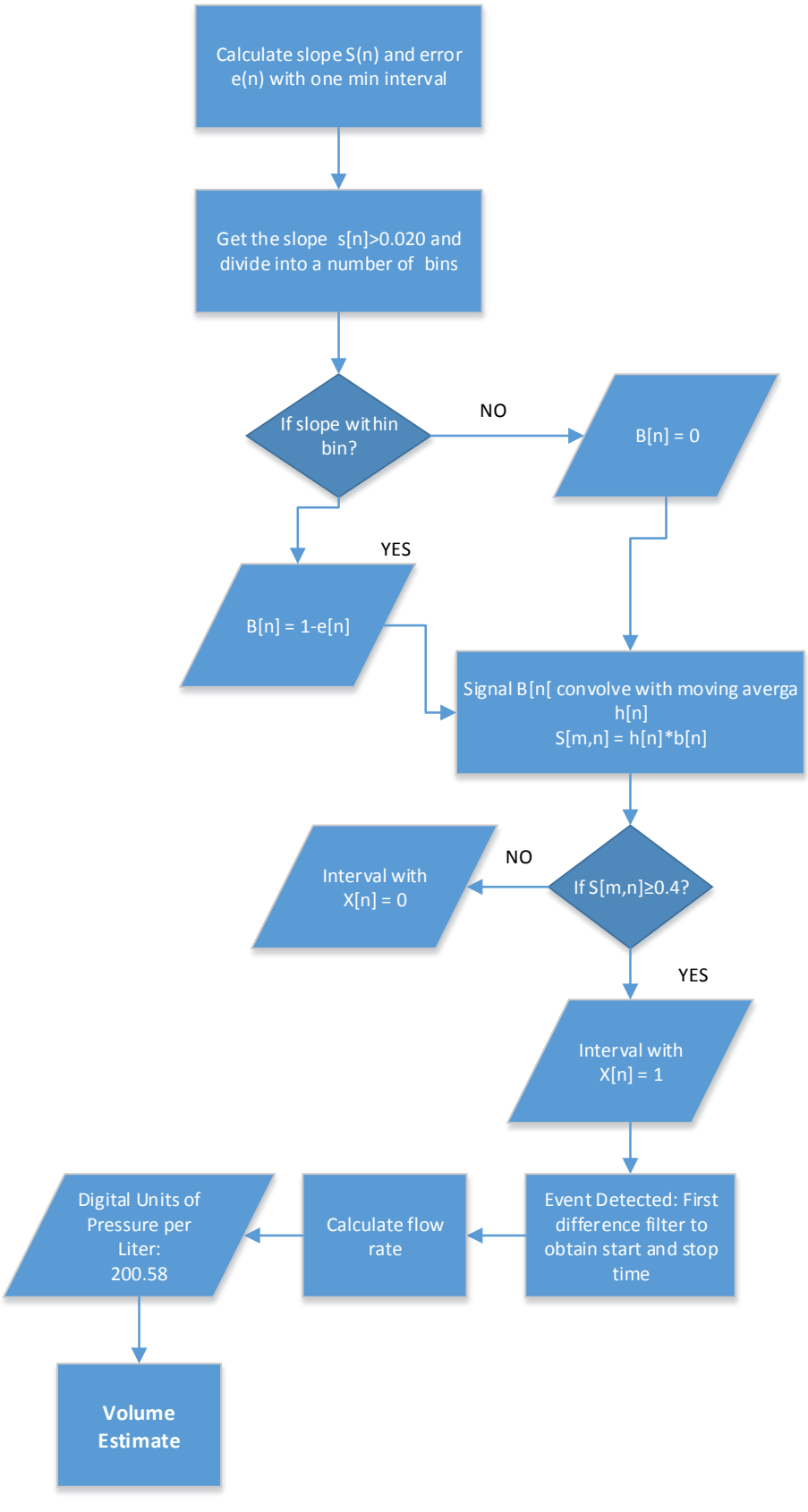
| **Test Case Name** | Unit Test – Multiple events per day | | | | **Test ID#:** | LSF20 -02 |
|---|---|---|---|---|---|---|
| **Description:** | Check if the algorithm is able to detect events that happen more than once a day. | | | | | |
| **Setup:** | 1. Pour five litters of water in different LifeStraws. Record time.<br>2. After two hours drain the water sitting at the bottom storage tank. Record time and volume.<br>3. Pour one more liter of water o amount water and drain water next day.<br>4. Repeat steps 1 through 3 for four, three, two, and one liters of water. | | | | | |
| **LifeStraw ID** | **Volume In** | **Time In** | **Volume Out** | **Time Out** | **Pressure** | **Comments** |
| LSF20_D47A15 | 5 | 5/22/15 14:12 | 3 | 5/22/15 16:13 | 8364.50 | |
| | 1 | 5/22/15 16:20 | 2.5 | 5/27/15 17:40 | 8365.56 | |
| | 4 | 5/27/15 17:48 | 4 | 6/9/17 16:30 | 8581.26 | |
| | 2 | 6/9/17 16:35 | 2 | 6/19/15 14:39 | 8251.07 | |
| | 1 | 6/19/15 14:42 | 1 | 6/19/15 16:00 | 7992.77 | |
| | 5 | 6/19/15 16:04 | 5 | 6/22/15 15:39 | 8814.00 | |
| | 4 | 6/22/15 15:42 | 4 | 6/23/15 16:50 | 8608.21 | |
| | 3 | 6/23/15 16:53 | 3 | 6/25/15 14:10 | 8369.16 | |
| | 2 | 6/25/15 14:12 | 2 | 6/25/15 16:10 | 8176.17 | |
| | 5 | 6/25/15 16:13 | 5 | 6/29/15 15:25 | 8803.51 | |
| | 2 | 6/29/15 15:28 | 1.5 | 6/29/15 16:37 | 8148.08 | |
| | 2 | 6/29/15 16:40 | 3 | 6/30/15 16:30 | 8398.79 | |
| LSF20_D47D0D | 5 | 5/22/15 14:05 | 3.01 | 5/22/15 16:10 | 8714.78 | |
| | 1 | 5/22/15 16:15 | 2.5 | 5/27/15 17:35 | 8653.45 | |
| | 4 | 5/27/15 17:40 | 4 | 6/9/17 16:37 | 8860.19 | |
| | 2 | 6/9/17 16:42 | 2 | 6/19/15 14:27 | 8556.82 | |
| | 1 | 6/19/15 14:30 | 1 | 6/19/15 15:57 | 8316.10 | |
| | 5 | 6/19/15 16:00 | 5 | 6/22/15 15:27 | 9112.62 | |
| | 4 | 6/22/15 15:37 | 4 | 6/23/15 16:44 | 8810.99 | |
| | 3 | 6/23/15 16:48 | 3 | 6/25/15 13:57 | 8634.52 | |
| | 2 | 6/25/15 14:08 | 2 | 6/25/15 16:06 | 8456.24 | |
| | 5 | 6/25/15 16:10 | 5 | 6/29/15 15:20 | 9119.51 | |
| | 2 | 6/29/15 15:25 | 1.5 | 6/29/15 16:32 | 8420.07 | |
| | 3 | 6/29/15 16:35 | 3 | 6/30/15 16:20 | 8657.20 | |
| Overall test result | | | | | | |

Acceptance Test

| **Test Case Name** | Acceptance Test | | | **Test ID#:** | |
|---|---|---|---|---|---|
| **Description:** | The acceptance test verifies that the algorithm meets the requirements. | | | | |
| **Setup:** | 1. Manually review round 2 LifeStraw data.<br>2. Run algorithm on round 2 LifeStraw data.<br>3. Check correlation between manual review and output of the algorithm. | | | | |
| Start Time | End Time | Sensor Filter | Start Time | End Time | Sensor Filter |
| 1432196692 | 1434538512 | LSF20_A63484 | 1432196692 | 1434538512 | LSF20_A63484 |
| 1432197539 | 1434540292 | LSF20_A635AD | 1432197539 | 1434540292 | LSF20_A635AD |
| 1431512551 | 1436016851 | LSF20_A5F4AE | 1431512551 | 1436016851 | LSF20_A5F4AE |
| 1431513060 | 1435931574 | LSF20_A60472 | 1431513060 | 1435931574 | LSF20_A60472 |
| 1431504481 | 1435931669 | LSF20_A5FE4F | 1431504481 | 1435931669 | LSF20_A5FE4F |
| 1431509086 | 1435933829 | LSF20_A61459 | 1431509086 | 1435933829 | LSF20_A61459 |
| 1431520577 | 1435929509 | LSF20_A6247B | 1431520577 | 1435929509 | LSF20_A6247B |
| 1431513311 | 1435922208 | LSF20_A621B3 | 1431513311 | 1435922208 | LSF20_A621B3 |
| 1431519808 | 1435920912 | LSF20_A61172 | 1431519808 | 1435920912 | LSF20_A61172 |
| 1431516119 | 1435918942 | LSF20_A6090A | 1431516119 | 1435918942 | LSF20_A6090A |
| 1431518227 | 1435918977 | LSF20_A63576 | 1431518227 | 1435918977 | LSF20_A63576 |
| 1431513596 | 1435921802 | LSF20_A60B44 | 1431513596 | 1435921802 | LSF20_A60B44 |
| Overall test result | | | | | |

# APPENDIX C: ALGORITHM FLOWCHART




Calculate slope S(n) and error e(n) with one min interval
Get the slope s[n]>0.020 and divide into a number of bins
If slope within bin?
NO
B[n] = 0
YES
B[n] = 1-e[n]
Signal B[n[ convolve with moving averga h[n]
S[m,n] = h[n]*b[n]
If S[m,n]≥0.4?
NO
Interval with X[n] = 0
YES
Interval with X[n] = 1
Event Detected: First difference filter to obtain start and stop time
Calculate flow rate
Digital Units of Pressure per Liter: 200.58
Volume Estimate
OUTPUT